Conditional Tabular Generative Adversarial Net for Enhancing Ensemble Classifiers in Sepsis Diagnosis

Ahmed Alfakeeh[a], Mhd Saeed Sharif[b], Abin Daniel Zorto[b] and Thiago Pillonetto[b]

[a]Research and Consultation Institute, King AbdulAziz University, Jeddah, SA.
[b]School of Architecture Computing and Engineering (ACE), UEL, University Way, London, E16 2RD, UK.

ARTICLE INFO

Keywords:
Sepsis
Machine Learning
Ensemble Classifier
Risk Factors
Deep Learning

ABSTRACT

Antibiotic-resistant bacteria have proliferated at an alarming rate as a result of the extensive use of antibiotics and the paucity of new medication research. The possibility that an antibiotic-resistant bacterial infection would progress to sepsis is one of the major collateral problems affecting people with this condition. 31,000 lives were lost due to sepsis in England with costs about two billion pounds annually. This research aims to develop and evaluate several classification approaches to improve predicting sepsis and reduce the tendency of underdiagnosis in computer-aided predictive tools. This research employs medical data sets for patients diagnosed with sepsis, it analyses the efficacy of ensemble machine learning techniques compared to non-ensemble machine learning techniques and the significance of data balancing and Conditional Tabular Generative Adversarial Nets for data augmentation in producing reliable diagnosis. The average F Score obtained by the non-ensemble models trained in this paper is 0.83 compared to the ensemble techniques average of 0.94. Non-ensemble techniques, such as Decision Tree, achieved an F score of 0.90, an AUC of 0.90 and an accuracy of 90%. Histogram-based Gradient Boosting Classification Tree achieved an F score of 0.96, an AUC of 0.96 and an accuracy of 95%, surpassing the other models tested. Additionally, when compared to the current state of the art sepsis prediction models, the models developed in this study demonstrated higher average performance in all metrics, indicating reduced bias and improved robustness through data balancing and Conditional Tabular Generative Adversarial Nets for data augmentation. The study revealed that data balancing and augmentation on the ensemble machine learning algorithms boost the efficacy of clinical predictive models and can help clinics decide which data types are most important when examining patients and diagnosing sepsis early through intelligent human-machine interface.

1. Introduction

Sepsis is a severe illness which is developed when the human body's reaction to a septicity leads to tissue damage and organ failure. For prompt and efficient treatment of sepsis, early detection is essential, since the mortality rate rises considerably with delayed diagnosis [1]. However, sepsis may be difficult to diagnose due to its broad and often mild symptoms and comorbidities [1]. Traditionally, sepsis has been diagnosed by clinical evaluation, laboratory testing, and imaging investigations. Research has been done in monitoring patients with sepsis using wearable sensor monitors in low- and middle-income countries [2]. Despite the fact that these techniques may give useful information, they may not always be adequate to provide an accurate diagnosis [3]. By examining a higher number of characteristics and using the power of data-driven decision-making, machine learning techniques, such as ensemble classifiers, have the potential to increase the accuracy of sepsis diagnosis [4]. Ensemble classifiers combine the predictions of numerous separate classifiers to provide a more accurate and dependable forecast [5]. Nonetheless, an imbalance in the class distribution in the data might impair the performance of ensemble classifiers [6]. Data balancing strategies, such as oversampling and under sampling [7], modify the number of samples in each class to enhance the classifier's capacity to learn from the data [8]. This research work will discuss the preparation of raw data, the generation of training and testing data, as well as the implementation, training, and visualization of a sepsis prediction model based on various methodologies.

This work is organized according to the following sections: Section 2 will analyze the related literature review on sepsis, its risk factors, and biomarkers. In addition, research on ensemble classifiers in the medical area will be examined. In Section 3, the utilized data set, its modifications, and its limits and limitations will be addressed in more depth. We provide details of the employed machine learning strategies to solve the classification issue and describe the models' architecture. In Section 4, the results and comments will be dissected and analyzed to offer a fuller picture of the findings of the research. In Section 5, based on the study's results, a variety of conclusions and recommendations will be presented.

2. Related Work

Several research studies have investigated the use of machine learning techniques, especially ensemble classifiers, in the diagnosis of sepsis. For instance, Fleuren et al. [9] conducted a comprehensive assessment of machine learning algorithms for sepsis detection and discovered that ensemble classifiers performed the best among the methods evaluated.

Several variables may influence the efficacy of machine learning approaches for sepsis detection, including the

Nomenclature

α	Intercept of linear equation	HGBC	Histogram Gradient Boosting Classifier
β	Gradient of linear equation	HR	Heart Rate
X	Independent variable	ICULOS	Intensive Care Unit Length of Stay
x_i	Distance of the i th instance	KNN	K Nearest Neighbors
Y	Binary target variable	LDA	Linear Discriminant Analysis
θ_i	Class of the i th instance	LR	Logistic Regression
ADA	AdaBoost Decision Tree	MLP	Multilayer Perceptron
AUC	Area Under Curve	NSGA-II	Non-Dominated Sorting Genetic Algorithm II
BC	Bagging Classifier	PTT	Partial Thromboplastin Time
BUN	Blood Urea Nitrogen	QDA	Quadratic Discriminant Analysis
CERF	Cox Enhanced Random Forest	Resp	Respiratory rate
DBP	Diastolic Blood pressure	RFC	Random Forest Classifier
DT	Decision Tree	RMSE	Root Mean Square Error
ETC	Extra Trees Classifier	SC	Stacked Classifier
GAN	Generative Adversarial Network	SVC	Support Vector Classifier
GBC	Gradient Boosting Classifier	SVM	Support Vector Machine
Hct	Hematocrit	VC	Voting Classifier
Hgb	Hemoglobin	WBC	White Blood Cell Count

amount of data used for training, the model's complexity, and the presence of noise or missing values in the data. Data balancing strategies, such as oversampling and under-sampling, have been suggested as a means of addressing class imbalance and enhancing the performance of machine learning systems for sepsis detection [7].

Mohan et al. [10] examined data from individuals diagnosed with sepsis who were monitored from the time they were admitted until either they passed away or were discharged from the intensive care unit over a two-year period. Their purpose was to aid in the development of improved algorithms by providing observation that resulted in mortality from septic shock. Machine learning was utilized by Mao et al. [11] To develop a prediction model utilizing just six routinely assessed and monitored vital indicators in medical institutes.

2.1. Risk Factors of Septic Shock

Studies have not shown that demographic factors have a major role in septic shock diagnosis. Age, gender, and length of stay are the three most significant demographic variables included in the data. In the majority of instances, age may be utilized as a significant predictor of sepsis risk. [12].

2.2. Biomarkers of Septic Shock

There have been several studies that have investigated the use of biomarkers for the diagnosis and prognosis of septic shock. For example, Lu et al. [13] developed a predictive model that used a combination of biomarker parameters to predict the risk of death in patients with septic shock. The scientists showed that the model had excellent discrimination and calibration and may be used to identify trauma patients at high risk for sepsis. Dellinger et al. [14] identified several biomarkers that have been proposed as indicators of septic shock, including procalcitonin, interleukin-6, and lactate. These biomarkers have been shown to be associated with the severity and prognosis of septic shock and may be useful for identifying patients at high risk of developing the condition.

Other studies have investigated the use of biomarkers in combination with clinical and laboratory parameters to improve the accuracy of septic shock diagnosis. To aid in the diagnosis of sepsis, researchers have developed a Lateral Flow Solid-Phase RPA for Sepsis-Related Pathogen Detection [15]. Quantitative identification of lactate using optical spectroscopy to help in continuous monitoring of serum lactate levels as a precondition for sepsis-prone patients requiring intensive care [16].

2.3. Ensemble Classifiers

Ensemble classifiers are classifiers which create a collection of hypotheses before combining them through weighted or unweighted voting [17]. The outcome of merging the separate selections is an improvement in overall performance and a more precise categorization [18].

There are three issues that diminish the performance of single classifiers: statistical, computational, and representational; These issues are handled by merging the findings and obtaining a better approximation [17].

The computational issue arises when the classification algorithm employs local optimization approaches that might

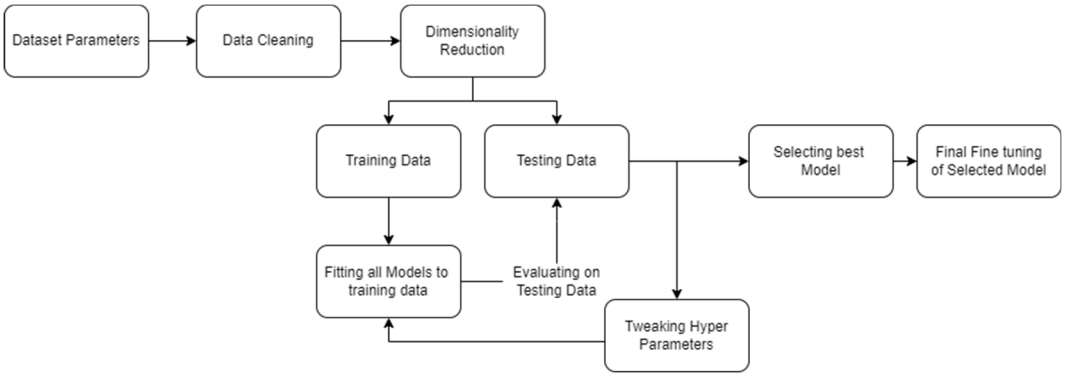

Figure 1: The developed approach for sepsis analysis.

get stalled at local minima (optima), preventing the process from discovering the optimal hypothesis [18].

2.4. Ensemble Classifiers in medicine

Lavanya and Rani [19] created a bagging-based ensemble classifier that was constructed from a collection of decision trees to increase the prediction accuracy of breast cancer detection. For the diagnosis of cardiac autonomic neuropathy, Kelarev et al. [20] utilized ensemble classification, and notably the Random Forest (RF), to produce a model with better abilities in prediction than those built on single classifiers.

For the purpose of predicting cancer survival, Gupta et al. [21] developed three models, each consisting of 400 SVM ensembles. The research found that using ensemble classifiers might improve prediction over traditional techniques [21]. Yao et al. [22] introduced a Random Forests- based ensemble classification method for predicting protein- protein interaction (PPI) networks.

2.5. Conditional Tabular Generative Adversarial Networks

Data generation plays a crucial role in various domains, including computer vision, natural language processing, and healthcare. Traditional approaches often rely on hand-crafted rules or statistical methods, which may not capture the complex underlying patterns of the data. Conditional Generative Adversarial Networks (cGANs) offer a promising solution by utilizing deep learning techniques to generate synthetic data that possesses desired characteristics [23].

Conditional Tabular Generative Adversarial Nets (CT-GAN) is a powerful technique in the field of generative adversarial networks (GANs) that specifically focuses on generating synthetic tabular data [24]. GANs have gained significant attention in recent years for their ability to generate realistic data that closely resembles the distribution of the training data. However, traditional GANs are not well- suited for tabular data generation due to the structured nature of such data. CTGAN addresses this limitation by incorporating conditional generation, allowing users to specify the desired attributes or conditions of the synthetic data [25]. This enables CTGAN to generate synthetic tabular data that not only resembles the distribution of the training data but also follows specific attributes or conditions set by the user [25]. This makes CTGAN a more suitable option for generating tabular data compared to traditional GANs. With the ability to generate realistic and customizable synthetic data, CTGAN opens up possibilities for various applications such as data augmentation, privacy preservation, and data analysis.

3. Materials and Methods

The proposed medical approach for sepsis analysis is illustrated in Figure 1. The acquired data sets go through the cleaning stage, where the missing parameters are identified, and missing data points are rectified. Following the dimensionality reduction, the data is split into training and testing data sets where several approaches will be evaluated. Different experiments have been performed to achieve the best approach structure that can generate the best performance.

3.1. Non-Ensemble Machine Learning Algorithms

3.1.1. Multinomial Logistic Regression

Multinomial regression is a variant of the binary regression model, in which both use logit analysis or logistic regression (LR) to get their conclusions. Logit analysis is a complement to linear regression and is especially beneficial when the response is a categorical variable.

For a binary target variable Y and an Independent Variable X, consider the following:

let: $\pi(x) = p(Y = 1|X=x) = 1-p(Y = 0|X=x)$, The logit of this probability may be expressed in linear form using the logistic regression model.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha+\beta x, \quad \text{with odds} = \exp(\alpha+\beta x), \tag{1}$$

The value of β is determined by the gradient of the S-shaped curve of $\pi(x)$. The curve is rising when β is positive, while the curve is descending when β is negative. The gradient's strength is inversely proportional to the strength of β [26].

3.1.2. Support Vector Machine for classification

To classify data, SVMs seek the hyperplane in a high-dimensional space that most clearly divides the classes [27]. Support vectors are the locations that are closest to the hyperplane, and the distance between the support vectors and the hyperplane is known as the margin. [27].

SVMs are particularly effective in cases where the number of dimensions is greater than that of the samples [27]. With the help of the hyperplane, the data may be projected into a lower-dimensional space, where the SVM can locate a separation border that was previously inaccessible [27]. The usage of support vector machines (SVMs) has spread across several fields, from text classification to picture classification to bioinformatics [27].

3.1.3. Multilayer Perceptron

An MLP is a neural network with numerous layers of linked "neurons," which are computational elements that take in data, analyze it, and output a result [28]). Each neuron in the MLP's levels gets input from all the neurons in the layer below it and sends its output to all the neurons in the layer above it because the MLP's layers are completely linked [28].

MLPs are often used for supervised learning tasks like classification and regression [29]. As part of their training, MLPs use optimization algorithms like stochastic gradient descent to fine-tune the weights of the connections between neurons in order to reduce the error between the expected and actual output [29]. Multiple-layer perceptrons, or MLPs, have been put to use in several fields, such as computer vision, NLP, and robotics [29].

3.1.4. Quadratic Discriminant Analysis (QDA)

QDA is based on another technique known as Linear Discriminant Analysis (LDA), which is based on the assumptions that the data is normally distributed and that the classes have identical covariance matrices [30]. Different class covariance matrices are acceptable in QDA, which may sometimes lead to better performance [30].

The purpose of QDA is to discover the decision boundary that optimally divides the classes based on their means and covariances [30]. The quadratic discriminant function, which is a function of the sample features and the class means and covariances, determines the quadratic decision boundary, as opposed to the linear decision boundary used in LDA [30]. QDA has been employed in a broad variety of applications, including text classification, picture classification, and predictive modelling [30].

3.1.5. Nearest Neighbor Classification

In a collection of n pairings where (x_1, θ_2) ,⋯,(x_1, θ_2) is predetermined, x_i takes values in an X metric space where d is defined, and θ_i takes values in the $\{1, 2, \cdots, M\}$ set. Every θ_i is regarded as the indication of the class that the i_{th} instance is a member of, and each x_i indicates the outcome of a set of tests conducted on the individual.

Given a new pair, (x, θ) in which only the measurement x may be observed, and it is wanted to estimate θ using the knowledge included in the collection of correctly identified points. If

$$\min\ d(x_I, x) = d\ (x_n, x)\ \ i = 1, 2, \cdots, n. \tag{2}$$

we will call

$$x'_{\approx} \varepsilon \{ x_1, x_2, \cdots, x_n \} \tag{3}$$

nearest neighbor to x

x is determined to belong to the category θ'_n of its nearest neighbor x'_n. If $\theta'_n \neq \theta$. an error has occurred. Only the nearest neighbors classification is used by the NN rule. The remaining $n-1$ classifications θ_i are disregarded.

Table 1
Non-Ensemble Model Parameters

Model	Hyperparameters
LR	-
SVC	-
MLP	-
QDA	-
KNN	-
DT	-

3.1.6. Decision Tree

A decision tree is a tree constructed using training data, where each leaf node denotes a label of a class and each internal node denotes a feature of the data. The classification is based on the feature values and the class labels of the training data. Decision trees are a popular machine learning method due to their interpretability and the ease with which they can be implemented [31]

3.2. Non ensemble model parameters

Table 1 illustrates the hyper-parameter information for the non-ensemble models in which we can see there have been no changes from the default parameters.

3.3. Ensemble Machine Learning Algorithms

3.3.1. Random Forest

A random forest is a kind of ensemble machine-learning technique in which numerous decision trees work together to produce an outcome that is the average of the classes produced by the individual trees. [32]. The individual decision trees are trained on different parts of the training set and use a random subset of the features to make predictions, resulting in a diverse set of trees that are able to capture different patterns in the data [32]. The use of multiple trees allows the random forest to make more accurate predictions than any individual tree would be able to make on its own [32]. The algorithm's error rate is proportional to the classification strength of each tree and the correlation between any two trees. Reducing the number of randomly selected qualities affects both the strength of each tree and the connection across trees, but increasing the number of randomly selected factors has the opposite effect [32].

3.3.2. Extra Trees Classifier

Extra trees, or extremely randomized trees, is a variant of the random forest algorithm [33]. Like random forests, extra trees are an ensemble method that consists of multiple decision trees. However, the decision trees in an extra trees' classifier are trained using random thresholds for each feature, rather than using the best split found during the training process as in a standard decision tree [33]. This results in a greater diversity of trees in the ensemble, which can lead to improved generalization performance [33].

3.3.3. AdaBoost Decision Tree

AdaBoost works by iteratively training weak classifiers and giving more weight to the instances that were misclassified in the previous iterations [34]. Weak classifiers are typically decision trees with a single split, known as decision stumps and the final strong classifier is the weighted sum of the weak classifiers, with the weight of each weak classifier being proportional to its accuracy [34]. AdaBoost has been shown to be a powerful and effective method for improving the performance of decision trees, especially when dealing with imbalanced or noisy datasets [34].

3.3.4. Bagging Classifier

According to Breiman et al. [35], In the bagging machine learning ensemble approach, many models are trained on various randomly chosen portions of the dataset, and the models are then combined to create a prediction. Bagging is intended to lower the model's variance by training the individual models in parallel and then combining their predictions. This can lead to improved generalization performance, especially when the training data is noisy or has a high variance. Bagging can be applied to any machine learning algorithm, but it is particularly effective for decision tree- based models, which have a tendency to overfit the training data.

3.3.5. Gradient Boosting Classifier

The goal of gradient boosting is to sequentially add weak learners to the ensemble, in a way that corrects the mistakes of the previous models. This is done by fitting the new model to the residual errors of the previous model, rather than to the original response. The final model is the weighted sum of the individual trees, with the weight of each tree being determined by the loss function. Gradient boosting has been shown to be a powerful and effective method for improving decision tree-based model performance, and it has seen extensive usage. [36].

3.3.6. Histogram Gradient Boosting Classifier

This classifier uses histograms to approximate the leaf values of the trees in the ensemble, rather than using exact leaf values as in traditional gradient boosting. This allows histogram gradient boosting to handle categorical features and large datasets more efficiently than traditional gradient boosting. In addition, histogram gradient boosting is more resistant to overfitting and can achieve higher predictive accuracy with fewer trees. Histogram gradient boosting has been shown to be a fast and effective method for improving the performance of decision tree-based models, and has been used in a wide range of applications [37].

Table 2
Ensemble Model Parameters

Model	Hyperparameters
RFC	criterion = "entropy", max_features=10
ETC	-
ADA	-
BC	-
GBC	learning_rate = 1
HGBC	learning_rate = 1
SC	estimators = estimators
VC	estimators = estimators

3.3.7. Stacked Classifier

A stacked Classifier (SC) is a strategy for reducing the biases of estimators by merging them [38]. Specifically, the estimators' outputs are stacked and fed into a single estimator to produce a final prediction. Cross-validation is used to train this final estimator [38]. The estimators used in this classifier will be comprised of the ensemble classifiers used in this research with its final estimator being the logistic regressor model.

3.3.8. Voting Classifier

Using the results of many base classifiers, a voting classifier makes a combined prediction. [18]. The final prediction is produced either by majority vote or by averaging the predictions of the basic classifiers, which may be trained using various algorithms and/or trained on separate subsets of the training data. [18]. When the base classifiers are varied and have varying strengths, a voting classifier may be utilized to increase the performance of a single classifier in a straightforward and effective manner [18]. The estimators used in this classifier will be comprised of the ensemble classifiers used in this research with its final estimator being the logistic regressor model.

3.4. Ensemble model parameters

Table 2 illustrates the hyper-parameter information for the ensemble models.

3.5. Dataset

The MIMIC-III dataset is a large database containing detailed information on patient demographics, vital signs, medications, laboratory test results, and clinical notes, among other things [39]. The MIMIC-III dataset is widely used in research on critical care and has been used to develop machine learning models for a variety of tasks [39].

The sepsis MIMIC-III dataset is a subset of the MIMIC-III dataset that includes only patients with a diagnosis of sepsis [1]. The sepsis MIMIC-III dataset includes detailed information on the clinical course of the sepsis, including the timing and dosage of interventions, as well as the patient's outcomes [1]. The sepsis MIMIC-III dataset is often used

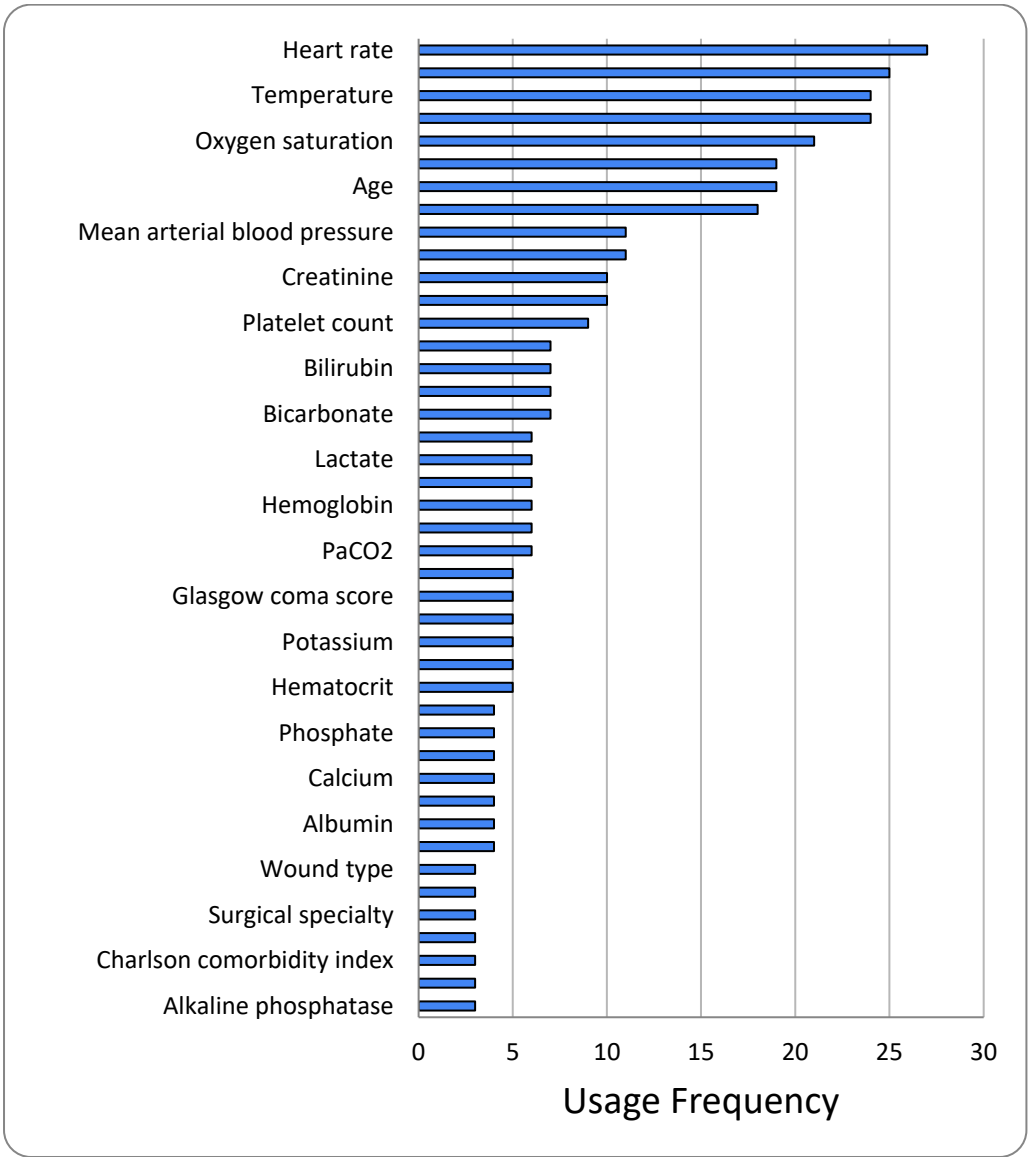

Figure 2: Dataset features and their usage frequency in current research. [9]

in research on sepsis and has been used to develop machine learning models for predicting patient outcomes and identifying sepsis in real-time [1].

Patients were monitored from the moment they entered the ICU, when t=0 until they were removed from the ICU or died. The database comprised 4,683 people aged 15 and above who had sepsis or severe sepsis. These patients had 8,696 admissions, 2,585 of which were due to septic shock. The data shown in Figure 3 illustrates the duration of time the patients examined in this data set were present, while Table 3 shows a summary describing the data set.

Table 3
Description of the Data Set

Data Set	A	B	Total
Patients	20,336	20,000	40,336
Septic patients	1,790	1,142	2,932
Prevalence	8.80%	5.70%	7.25%
Rows	739,663	684,508	1,424,171
Entries	5,536,849	4,950,064	10,486,913
Density of entries	20.60%	19.10%	19.85%

3.5.1. Dataset Limitations

The dataset is imbalanced with 2932 patients with a sepsis diagnosis whereas there are over 37000 patients without a sepsis diagnosis. A comprehensive analysis of the data set revealed that certain attributes are totally empty, indicating that if they are not eliminated, the training set will be misled or an improperly functioning model would be generated an example of this is shown in Figure 4.

3.5.2. Dataset Manipulation and Delimitation

This dataset contains 2932 diagnosed sepsis patients compared to 37404 patients without a diagnosis. This is resolved by augmenting the sepsis patient data by generating

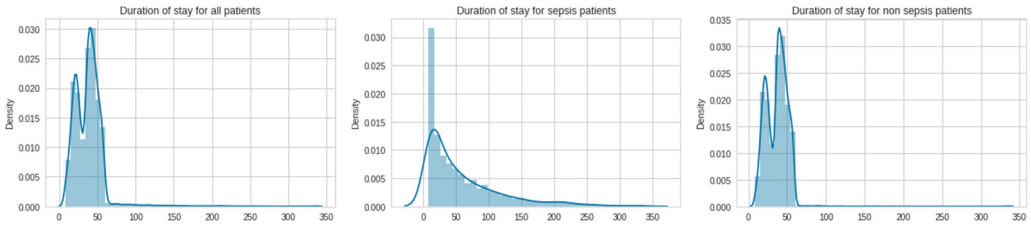

Figure 3: Duration of stay (All patients left, sepsis patients middle, non-sepsis patients right)

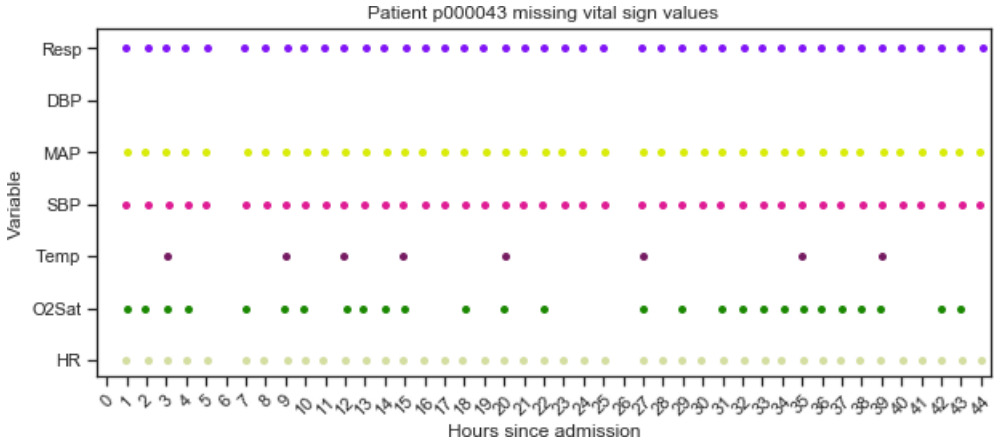

Figure 4: Missing vitals for a patient.

2068 sepsis patients and then taking the first 5000 non-diagnosed patients and ignoring the remaining 32404 to prevent the dataset from prioritizing non-diagnosed patients during training.

Researchers often encounter the difficulty of missing data. This dataset includes components with real number values, and missing data, which will be filled in using an interpolation function that substitutes NaN values with values that have no influence on the final result but optimize the model. The sum of all attributes will be used to calculate the fraction of missing data, and this parameter will be adjusted to generate the most effective models.

The possibility of removing attributes from the training process will also be considered based on their correlation to the target variable as well as their frequency of use in current research as shown in Figure 2

3.5.3. F score Recall and AUC for model selection

The F score is a class-balanced accuracy metric since it represents the weighted harmonic mean of precision and recall. When false negatives and false positives are important in the prediction process, the F1-score is utilized. Current research shows that most sepsis prediction models for this data set are more adept at predicting non-diagnosed patients than diagnosed patients [4]. This is due to unbalanced classes and the fact that most instances in the data are classified as non-sepsis patients leading the accuracy of non-sepsis predicted cases to dominate the overall accuracy measure.

Recall is an important metric for measuring a model's ability to detect positive samples in which the higher the recall, the more positive samples are detected. For the purpose of machine learning in clinical settings, it can be argued that true positives are more important than true negatives as an undetected true positive can lead to a fatality whereas an undetected true negative is not fatal.

AUC represents the area under the ROC (Receiver Operating Characteristic) curve, which plots the true positive rate against the false positive rate at different classification thresholds [40]

3.5.4. Methodology Comparison

The three works compared in this paper focus on predicting and diagnosing sepsis, but they differ in their approaches, methodologies, and evaluation metrics. While this research aims to improve sepsis prediction and reduce underdiagnosis through the use of machine learning algorithms. It evaluates ensemble and non-ensemble machine learning techniques, employs data balancing and augmentation through the use of CTGAN, and reports F score, AUC and accuracy as evaluation metrics. El-Rashidy et al., [41] proposes a multi- stage model for sepsis prediction that combines NSGA- II, artificial neural networks, and deep learning models. It utilizes NSGA-II and neural networks to extract the optimal feature subset from patient data. The model consists of a deep learning classification model and a multitask regression model to predict sepsis, onset time, and blood pressure. It uses the MIMIC III real-world dataset and reports accuracy,

Table 4
Attribute correlation to sepsis diagnosis

Variable	Correlation
ICULOS	0.39
Calcium	0.26
BUN	0.22
HR	0.21
Resp	0.20
Creatinine	0.18
Temp	0.17
Hgb	0.17
Fibrinogen	0.16
PTT	0.14
Bilirubin_total	0.14
Hct	0.12
HospAdmTime	0.11
WBC	0.10
DBP	0.10

Table 5
Average Model Performance Comparison

Correlation	F score	Accuracy	Recall	AUC
Top 15	0.89	88	0.89	0.89
All Included	**0.89**	**89**	**0.89**	**0.89**

specificity, sensitivity, AUC, and RMSE as evaluation metrics. Darwiche et al., [4] focuses on developing an improved method for predicting septic shock. It trains an ensemble classifier using the MIMIC-III database and incorporates the Cox Hazard model to obtain a risk score. The Random Forest ensemble classifier is trained using this score and other features. Specific evaluation metrics are not mentioned, but the predictive accuracy of the proposed CERF method is compared to existing methods. Overall, each study presents a unique approach to sepsis prediction and diagnosis, show- casing different techniques and evaluation criteria.

4. Results

4.1. Correlation of Sepsis factors

After quantitative analysis using the pandas Python library, we analyzed the dataset and produced Table 4 which shows us the 15 variables with the highest correlation to a sepsis diagnosis. These correlation values can give more insight into the type of data to be collected for processing in order to aid diagnosis [42]. Table 5 illustrates that the results attained by selecting the top 15 correlated attributes for training produces lower performance versus selecting for all attributes. Thus, for the training and tuning of the final selected model we used models trained on all attributes regardless of correlation. The missing values in the data are also filled with the mean value of each attribute so as to make the data more quantitatively meaningful.

Table 6
Non-Ensemble Model Performance Results

Model	F Score	Accuracy	Recall	AUC
LR	0.80	80	0.80	0.80
SVC	0.79	79	0.79	0.79
MLP	0.89	87	0.89	0.89
QDA	0.82	82	0.82	0.82
KNN	0.81	80	0.81	0.81
DT	**0.90**	**90**	**0.90**	**0.90**
Average	0.84	83	0.84	0.84

4.1.1. Machine Learning Model Evaluation and Performance Analysis

The code performs the training and testing of machine learning models to predict and evaluate sepsis. It uses a popular library called scikit-learn, which is widely used for machine learning in Python. The dataset is divided into two parts: a training set and a testing set. The training set is utilized in conjunction with 10-fold cross-validation to train the models. This approach enables a more efficient utilization of the available data, as all observations are utilized for both training and validation purposes [43]. Additionally, it is less susceptible to variations in the precise manner in which the data is partitioned, in comparison to alternative methods [44]. The testing set is used to evaluate the model's performance.

The code follows these steps:

1. The dataset is prepared and split into input features (such as patient information) and the target variable (whether a patient has sepsis or not).
2. using CTGAN the data is augmented to provide more data for training and testing.
3. A portion of the dataset is set aside for testing the trained models.
4. Different machine learning models, such as logistic regression, decision trees, and ensemble models, are trained using the training data. During the training process, the models are subjected to 10-fold cross-validation in order to mitigate potential sources of unreliability and bias. This approach aims to enhance the model's ability to discern meaningful patterns from the available data and generate dependable predictions.
5. After training the models, their performance is evaluated using various metrics, including accuracy (how often the model is correct), sensitivity (how well the model detects positive cases), specificity (how well the model detects negative cases), and F-score (a combined measure of precision and recall). These metrics help assess how well the models can predict sepsis.
6. The evaluation results, such as accuracy, sensitivity, and specificity, are recorded for further analysis.

4.2. Model Performance

Table 6 displays that the Decision Tree model, with an accuracy of 90%, an AUC of 0.90, and a F score of

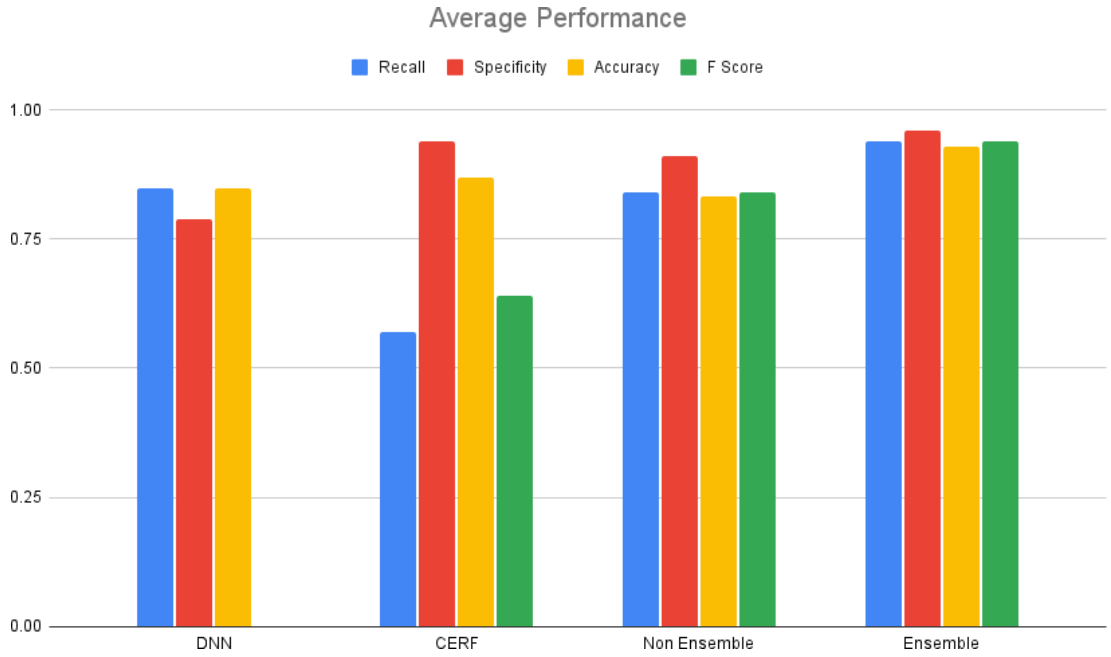

Figure 5: Average Performance

Table 7
Ensemble Model Performance Results

Model	F Score	Accuracy	Recall	AUC
RFC	0.95	94	0.95	0.95
ETC	0.93	92	0.93	0.92
ADA	0.94	93	0.94	0.94
BC	0.93	93	0.93	0.93
GBC	0.93	93	0.93	0.93
HGBC	0.92	92	0.92	0.92
SC	**0.95**	**95**	**0.95**	**0.95**
VC	0.94	94	0.94	0.94
Average	0.94	93	0.94	0.94

0.90, is the best-performing model among the non-ensemble strategies.

With a F score of 0.95, an AUC of 0.95, and an accuracy of 95%, Table 7 demonstrates that the stacking classifier model is the best-performing model among the ensemble strategies.

4.3. Further Testing and Tuning

The results of further testing and tuning for the histogram-based Gradient Boosting Classification Tree model are presented in Tables 8 and 9. The tables show the performance metrics, including F score, accuracy, recall, and AUC, for different values of the L-rate and regularization (L2) parameters, respectively.

Table 9 shows the best-performing model from the ensemble techniques is the Histogram-based Gradient Boosting Classification Tree model with an F Score, Accuracy, Recall and AUC of 0.96, 95, 0.96 and 0.96

Table 8
Histogram-based Gradient Boosting Classification Tree L-Rate Tuning

L-Rate	F Score	Accuracy	Recall	AUC
0.10	0.95	95	0.95	0.95
0.20	0.95	95	0.95	0.95
0.30	0.95	95	0.95	0.95
0.40	0.95	95	0.95	0.95
0.50	0.95	95	0.95	0.95
0.60	0.95	95	0.95	0.95
0.70	0.94	94	0.94	0.94
0.80	0.95	93	0.95	0.95
0.90	0.93	93	0.93	0.93

respectively.

Figure 6 shows the confusion matrix for the selected model in which we can see that the model is accurate at predicting sepsis and non-sepsis patients.

The findings suggest that there is a possibility of enhancing the performance of the model by the modification of these hyperparameters. Additionally, it may be beneficial to prioritize minimizing instances of non-detection of sepsis patients, even if it leads to an increase in the diagnosis of sepsis patients, as failure to do so could have severe consequences. These findings emphasize the importance of thorough testing and tuning of model hyperparameters to optimize the performance of the histogram-based Gradient

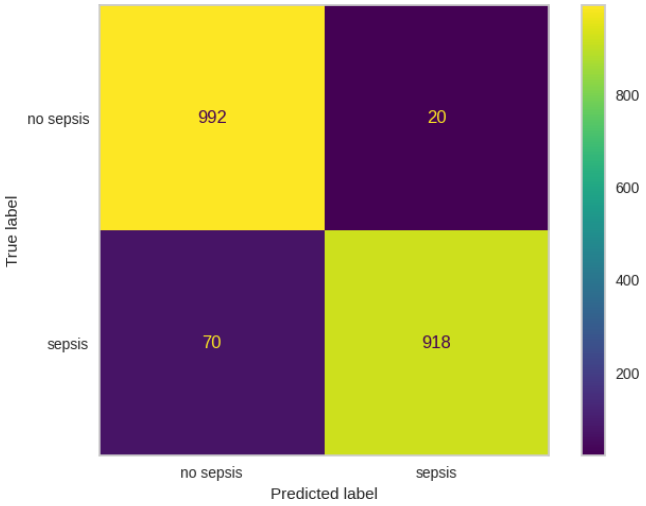

Figure 6: Confusion Matrix

Table 9
Histogram-based Gradient Boosting Classification Tree regularisation Tuning

L2	F Score	Accuracy	Recall	AUC
1	0.96	95	0.96	0.96
2	0.95	95	0.95	0.95
3	0.95	95	0.95	0.95
4	0.95	95	0.95	0.95
5	0.95	95	0.95	0.95
6	0.95	95	0.95	0.95
7	0.96	95	0.96	0.96
8	0.96	95	0.96	0.96
9	0.95	95	0.95	0.95

Boosting Classification Tree model. Further exploration and fine-tuning of these parameters can lead to improved accuracy, F score, recall, and AUC, thus enhancing the model's predictive capabilities and overall effectiveness.

4.4. Average Performance Comparison

Figure 5 illustrates the average performance of the models created in this paper compared to the CERF models created by Darwiche, [4] and the Ensemble DNN models by El-Rashidy [41]. The models in this paper produce higher average F scores, AUC and Recall showing the ability of the machine learning models to produce more robust pre- dictions with a lower risk of bias in prediction. This paper's strengths lie in its robust performance, potential novelty in reintroducing machine learning techniques, and rigorous experimental evaluation. However, potential weaknesses include the need for further generalizability testing on diverse datasets and real-world scenarios, limited comparisons with existing state-of-the-art methods, and a potential lack of interpretability in the proposed models.

The DT and HGBC models are the only models in Table 10 comparison using CTGAN for data augmentation with

Table 10
Best Model Performance Comparison

Technique	F score	Accuracy	Recall	AUC
DT	0.90	90	0.90	0.90
HGBC	**0.96**	**95**	**0.96**	**0.96**
NSGA-II[41]	-	91	0.92	0.91
CERF[4]	0.9	95	0.89	-

the best performance is the HGBC model with 95% in Accuracy an F score of 0.96 a Recall of 0.96 and an AUC of 0.96. Based on these results the selected model for this paper is the HGBC model.

5. Conclusion

The developed ensemble machine learning-based algorithm holds substantial importance in the clinical sector. By achieving improved efficacy in predictive models, it addresses the critical need for accurate disease diagnosis and prognosis. This algorithm can potentially revolutionize medical practices by assisting clinicians in making more informed decisions and providing better patient care.

The research study highlights the necessity of employing generative data-balancing techniques such as CTGAN in the training process. Imbalanced datasets can lead to biased models and under-diagnosis of illnesses, which can have severe consequences in certain situations. By demonstrating the effectiveness of data balancing and augmentation, the research emphasizes the need for mitigating bias and ensuring accurate predictions in healthcare applications.

The HGBC model with 95% Accuracy, an F score of 0.96, a Recall of 0.96, and an AUC of 0.96 had the highest performance on the sepsis data. Based on these results, the selected model for this paper is the HGBC model, which combines multiple base classifiers to improve overall prediction performance. The findings provide valuable insights for

researchers and practitioners in selecting the most effective model for sepsis prediction. We suggest that future work should focus on gathering more data on risk factors to improve disease diagnosis. Additionally, parameter tuning is identified as a crucial step to enhance the effectiveness of the models. By exploring different datasets, processing techniques, and algorithms, the research encourages further validation and fine-tuning of predictive models in order to optimize their performance.

The research holds the potential to significantly impact clinical practice by providing an effective computer-aided medical prediction approach. The developed algorithm, coupled with intelligent human-machine interfaces, can aid clinicians in early disease detection and improve patient outcomes. The research lays the foundation for further advancements in computer-aided diagnostics and personalized medicine.

Acknowledgment

The author gratefully acknowledges the deanship of scientific research (DSR) technical and financial support, the Ministry of Education and King Abdulaziz University, under grant no. (IFPIP: 1018-611-1443).

References

[1] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, Richard S Hotchkiss, Mitchell M Levy, John C Marshall, Greg S Martin, Steven M Opal, Gordon D Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, February 2016.

[2] Shadi Ghiasi, Tingting Zhu, Ping Lu, Jannis Hagenah, Phan Nguyen Quoc Khanh, Nguyen Van Hao, Vital Consortium, Louise Thwaites, and David A Clifton. Sepsis mortality prediction using wearable monitoring in Low-Middle income countries. *Sensors*, 22(10), May 2022.

[3] Karen Nagalingam. Understanding sepsis. *Br. J. Nurs.*, 27(20):1168–1170, November 2018.

[4] Aiman Darwiche and Sumitra Mukherjee. Machine learning methods for septic shock prediction. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality*, AIVR 2018, pages 104–110, New York, NY, USA, November 2018. Association for Computing Machinery.

[5] L K Hansen and P Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, October 1990.

[6] N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, June 2002.

[7] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. ieeexplore.ieee.org, June 2008.

[8] Gustavo E A P A Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004.

[9] Lucas M Fleuren, Thomas L T Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand R J Girbes, Patrick Thoral, Ari Ercole, Mark Hoogendoorn, and Paul W G Elbers. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.*, 46(3):383–400, March 2020.

[10] Anant Mohan, Prajowl Shrestha, Randeep Guleria, Ravindra Mohan Pandey, and Naveet Wig. Development of a mortality prediction formula due to sepsis/severe sepsis in a medical intensive care unit. *Lung India*, 32(4):313–319, July 2015.

[11] Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, Yifan Zhou, and Ritankar Das. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*, 8(1):e017833, January 2018.

[12] Soufiane Chami and Kouhyar Tavakolian. Early prediction of sepsis from clinical data using single Light-GBM model. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4. ieeexplore.ieee.org, September 2019.

[13] Hong-Xiang Lu, Juan Du, Da-Lin Wen, Jian-Hui Sun, Min-Jia Chen, An-Qiang Zhang, and Jian-Xin Jiang. Development and validation of a novel predictive score for sepsis risk among trauma patients. *World J. Emerg. Surg.*, 14:11, March 2019.

[14] R P Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, Tiffany M Osborn, Mark E Nunnally, Sean R Townsend, Konrad Reinhart, Ruth M Kleinpell, Derek C Angus, Clifford S Deutschman, Flavia R Machado, Gordon D Rubenfeld, Steven Webb, Richard J Beale, Jean-Louis Vincent, Rui Moreno, and Surviving Sepsis Campaign Guidelines Committee including The Pediatric Subgroup. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive Care Med.*, 39(2):165–228, February 2013.

[15] Alice Jane Heeroma and Christopher Gwenin. Development of solid-phase rpa on a lateral flow device for the detection of pathogens related to sepsis. *Sensors*, 20(15):4182, 2020.

[16] Karthik Budidha, Mohammad Mamouei, Nystha Baishya, Meha Qassem, Pankaj Vadgama, and Panayiotis A. Kyriacou. Identification and quantitative determination of lactate using optical spectroscopy—towards a noninvasive tool for early recognition of sepsis. *Sensors*, 20(18), 2020.

[17] Giorgio Valentini and Francesco Masulli. Ensembles of learning machines. In *Neural Nets*, pages 3–20. Springer Berlin Heidelberg, 2002.

[18] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer Berlin Heidelberg, 2000.

[19] D Lavanya and K Usha Rani. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, 2(1):17–24, 2012.

[20] Andrei Kelarev, Richard Dazeley, Andrew Stranieri, John Yearwood, and Herbert Jelinek. Detection of CAN by ensemble classifiers based on ripple down rules. In *Knowledge Management and Acquisition for Intelligent Systems*, pages 147–159. Springer Berlin Heidelberg, 2012.

[21] Sunil Gupta, Truyen Tran, Wei Luo, Dinh Phung, Richard Lee Kennedy, Adam Broad, David Campbell, David Kipp, Madhu Singh, Mustafa Khasraw, Leigh Matheson, David M Ashley, and Svetha Venkatesh. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open*, 4(3):e004007, March 2014.

[22] Jianzhuang Yao, Hong Guo, and Xiaohan Yang. PPCM: Combing multiple classifiers to improve Protein-Protein interaction prediction. *Int. J. Genomics Proteomics*, 2015:608042, October 2015.

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[24] Lei Xu. *Synthesizing Tabular Data Using Conditional GAN*. Jan 2020.

[25] Basim Ahmad Alabsi, Mohammed Anbar, and Shaza Dawood Ahmed Rihan. Conditional tabular generative adversarial based intrusion detection system for detecting ddos and dos attacks on the internet of things networks. *Sensors*, 23(12):5644, Jun 2023.

[26] Abdalla M El-Habil. An application on multinomial logistic regression model. *PJSOR*, pages 271–291, March 2012.

[27] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[28] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
[29] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
[30] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.
[31] L Breiman, J Friedman, R Olshen, and C Stone. Cart. *Classification and Regression Trees*, 1984.
[32] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.
[33] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, April 2006.
[34] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1):119–139, August 1997.
[35] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, August 1996.
[36] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29(5):1189–1232, 2001.
[37] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, 30, 2017.
[38] David H Wolpert. Stacked generalization. *Neural Netw.*, 5(2):241–259, January 1992.
[39] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, May 2016.
[40] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
[41] Nora El-Rashidy, Tamer Abuhmed, Louai Alarabi, Hazem M El-Bakry, Samir Abdelrazek, Farman Ali, and Shaker El-Sappagh. Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. *Neural Comput. Appl.*, 34(5):3603–3632, March 2022.
[42] Venish Suthar, Vinay Vakharia, Vivek K Patel, and Milind Shah. Detection of compound faults in ball bearings using multiscale-singan, heat transfer search optimization, and extreme learning machine. *Machines*, 11(1):29, 2022.
[43] V Malathi, MP Gopinath, Manoj Kumar, Shashi Bhushan, Sujith Jayaprakash, et al. Enhancing the paddy disease classification by using cross-validation strategy for artificial neural network over baseline classifiers. *Journal of Sensors*, 2023, 2023.
[44] Vinay Vakharia, Milind Shah, Pranav Nair, Himanshu Borade, Pankaj Sahlot, and Vishal Wankhede. Estimation of lithium-ion battery discharge capacity by integrating optimized explainable-ai and stacked lstm model. *Batteries*, 9(2):125, 2023.