Interpretable classification of imbalanced medical data with greedy decision trees based on random forests

Bachelor thesis

Jan Büttner
411563

November 12, 2024

Supervisor: Prof. Dr. Benjamin Blankertz
Prof. Dr. Klaus-Robert Müller

Technische Universität Berlin
School of Electrical Engineering and Computer Science
Institute of Software Engineering and Theoretical Computer Science

Kurzfassung

In dieser Arbeit wird die Nutzung von Greedy-Entscheidungsbäumen auf Basis von Zufallswäldern zur Klassifikation von imbalancierten medizinischen Daten untersucht. Aufgrund der häufigen Klassenungleichgewichte in medizinischen Datensätzen stehen herkömmliche Algorithmen vor der Herausforderung, Minderheitsklassen, die oft kranke Patienten darstellen, präzise zu klassifizieren.

Die Arbeit untersucht die Anwendung von Greedy-Entscheidungsbäumen auf Basis von Zufallswäldern im Vergleich zu Zufallswäldern und einfachen Entscheidungsbäumen, um zu entscheiden, ob sich die zusätzliche Rechenzeit für die Erstellung der Greedy-Entscheidungsbäume lohnt. Ziel ist es, festzustellen, ob sich dadurch eine bessere Klassifikation erzielen lässt und ob sich Greedy-Entscheidungsbäume gegenüber einfachen Entscheidungsbäumen als vorteilhaft erweisen, um eine erklärbare Klassifikation zu erhalten.

Die Methoden wurden auf zwei medizinischen Datensätzen angewendet: dem MIMIC-III-Datensatz zur Vorhersage von Mortalität und einem Datensatz zur Analyse postoperativer Delirium-Raten. Die Effektivität des Ansatzes wurde durch Evaluation mit Metriken wie der ROC-Kurve und dem AUC-Wert im Vergleich zu traditionellen Methoden bewertet. Die Ergebnisse zeigen, dass auf unausgewogenen medizinischen Datensätzen Greedy-Entscheidungsbäume nicht generell besser klassifizieren als einfache Entscheidungsbäume und dass eine vorherige Analyse der Leistung des Greedy-Entscheidungsbaums auf den Daten stattfinden muss.

Abstract

This paper investigates the use of greedy decision trees based on random forests to classify imbalanced medical data. Due to the frequent class imbalances in medical datasets, conventional algorithms face the challenge of accurately classifying minority classes, which often represent sick patients.

The thesis investigates the application of greedy decision trees based on random forests in comparison to random forests and simple decision trees to decide whether the additional computing time for creating greedy decision trees is worthwhile. The aim is to determine whether a better classification can be achieved and whether greedy decision trees prove to be advantageous compared to simple decision trees in order to obtain an explainable classification.

The methods were applied to two medical datasets: the MIMIC-III dataset for predicting mortality and a dataset for analyzing post-operative delirium rates. The approach's effectiveness was assessed by evaluation with metrics such as the ROC curve and the AUC value compared to traditional methods. The results show that on imbalanced medical datasets, greedy decision trees do not generally classify better than simple decision trees and that prior analysis of the performance of the greedy decision tree on the data needs to take place.

Contents

List of Figures

List of Tables

1 Introduction

In recent years, the multiplication of medical data has presented opportunities, such as data mining, to find correlations that were previously undiscovered and challenges like building methods that are interpretable and accurate at the same time [5]. Accurate and interpretable models for classifying medical data are crucial, as they can significantly impact the diagnosis and treatment planning that a doctor can prescribe and, therefore, change patient outcomes.

However, medical data sets often exhibit imbalanced class distributions, where some conditions or outcomes are under-represented compared to others. This imbalance poses a significant challenge for traditional machine-learning algorithms, which tend to be biased towards the majority class. This leads to poor performances of accurately classifying the minority class, which is often of greater clinical importance. The minority class is usually the one that contains the patients that were diagnosed with a disease [9, 14, 33].

To address this issue, researchers have increasingly turned to ensemble learning methods, such as Random forests, due to their ability to classify imbalanced data accurately despite the existing bias. They combine the predictions of multiple decision trees to improve classification accuracy [19]. Despite their superior predictive performance, random forests are often criticized for their black-box nature, as the complexity of the model makes it difficult to interpret how decisions were made [12]. In the medical field, interpretability is essential, as physicians and patients need to understand and trust the models they use to make informed decisions about patient care.

This thesis explores the use of greedy decision trees derived from random forests to classify imbalanced medical data. The greedy decision tree approach uses a decision forest which was trained on data that is then ensemble pruned [31]. The forest is merged into on model that is then organized into a tree structure. The tree can be easily interpreted through text or a figure of the tree [13, 29] without sacrificing a lot of the predictive performance provided by the ensemble model [31].
The objective of this study is to investigate the suitability of the proposed method for building interpretable classifiers in the context of medical data sets. In particular, the focus is on assessing the effectiveness of the method by addressing the challenges posed by imbalanced medical data and its potential for use in future studies aimed at developing more accurate diagnostic tests for patients.

I will lay out the scientific methods necessary for this thesis in Chapter 2. Chapter 3 shows the experiment's results, examining the effectiveness of this approach and traditional methods on various imbalanced medical data sets and the models trained on balanced data regarding classification accuracy and interpretability.
In Chapter 4, I will discuss my findings and their implications for using the forest-based tree on imbalanced medical data. In Chapter 5, I will conclude and present aspects of the forest-based tree based on imbalanced medical data that could be further investigated.

2 Methods

Physicians need to understand the reasoning behind a model's decision to provide the best possible service to their patients. Therefore a simple and understandable machine-learning model is critical in medicine. However, because most data in the medical field is imbalanced, different models are needed to address the challenges of learning from imbalanced data.
I will analyze the method introduced in Omer Sagis and Lior Rokach's paper [31] that builds a decision tree out of a decision forest that approximates the predictive power of the forest [31]. In the following section, I will provide the relevant scientific methods by examining decision trees, learning from imbalanced data, ensemble learning, and decision forests as an approach to tackle learning problems on imbalanced data.

2.1 Decision tree

A decision tree is a type of predictive model used in machine-learning that maps values of features in data to decisions about a target value. It resembles a flowchart, where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents a prediction or classification [23, 26].
In decision trees nodes represent decisions or tests on a feature. Branches represent the outcome of each decision. Leaves represent the final classification [23, 26]. An example of this can be seen in figure 2.1.
Decision trees are widely used because they are simple to interpret, can tackle both numerical and categorical problems, and don't require extensive data preprocessing. However, they can become complex and prone to overfitting [23, 1, 26]. Overfitting is when a machine-learning model sticks too closely to the data on which the model was trained and, therefore, fits new entries less accurately [37].

2.2 Random forest

A random forest is an ensemble learning method used in machine-learning. It is composed of multiple decision trees that all learn from the same data set to solve classification problems. Every tree in the forest makes a prediction that the forest then averages to make a definitive decision [3, 30].
Each tree is trained on a random subset of the data. Due to the randomness, overfitting is reduced, and the abstraction of new data is improved. The ensemble method increases classification accuracy and stability, often outperforming single decision trees [3, 30].
Random forests are favored because they have high accuracy and strong robustness against overfitting [3, 30, 19]. However, they may require more computational resources and can be harder to interpret than single decision trees [30, 29].

2.3 Learning from imbalanced data

Imbalanced data in the context of classification means that the amount of representatives of one class is higher (majority class) than the other classes (minority classes) [15].
This brings unique challenges for machine-learning; because the majority class is more represented, the learning method will develop a bias for this class, providing a poor prediction result for the minority classes.
In the domain of medical data, the minority class is usually the more important one, for there are many healthy people and significantly fewer sick people. Thus, methods are needed to resolve this problem such that sick patients can be predicted robustly with a high accuracy [33].
One common method to address class imbalance is to even out the number of class representatives. Two possible methods are: Producing more entries of the minority class by randomly over-sampling it or eliminating entries of the majority class by randomly under-sampling it. The first comes with the drawback of creating empty data by duplicating minority class entries [11]. The latter comes with the fact that a lot of data is unused, which results in losing prediction power [11, 20].
Another way to address learning problems from imbalanced data is to use different learning algorithms designed to combat the bias that comes with class imbalance [9]. One type of algorithm often used to learn from imbalanced data are ensemble methods, in many cases, the random forest [19]. On imbalanced data, a random forest performs well because the forest, unlike one tree, does not focus on one global maximum in the data but can focus on many aspects. So, it can combat the flaw single classifiers have on imbalanced data very well [19].
One disadvantage of the random forest and ensemble classification, in general, is their complexity and interpretability.

2.4 Complexity reduction of ensemble classification methods

2.4.1 Pruning functions

Ensemble methods use more memory and are more complex than single-classifier methods [24, 34]. However, more complexity is not always better. More complexity means a less interpretable model, more computation time for a decision, and, in some cases, poorer prediction results. When the estimators used by the ensemble methods are too similar, then the combined prediction of the ensemble method is less accurate than the prediction of an ensemble method that has fewer more diverse estimators [38].
A way to address the complexity of the random forest is pruning functions. Pruning functions reduce the complexity of ensemble methods by finding a subset of classifiers, in this case, a subset of decision trees that performs as well or better than the original set [38].
They are selected for the remaining subset by ranking the classifiers. The ranking method used to prune the forest during the creation of the forest-based tree applies a cross-validation-based search to reduce the error of predictions for the new subset of estimator [8, 16, 31].

2.4.2 Conjunction sets

A conjunction set is an ensemble-derived model classifier that consists of a set of rules. The ensemble-derived model aims to produce classifications swifter and more comprehensible while maintaining the

predictive performance of the initial ensemble [35].
It is built from a decision forest by pairwise conjoining the decisions of two non-conflicting paths of two decision trees from the forest [31].
This reduces the complexity of an ensemble classification model by creating a new single model from it with the accuracy performance score. To achieve this, each path of the trees is seen as a set of rules mapped to a vector that contains probabilities for each class in its cell. Applying a Cartesian Product can merge conjunctions to build a new conjunction that represents both of them.
The resulting conjunction set represents every possible outcome of the given decision forest [31].

2.5 Forest-based tree

The forest-based tree is a classification model that aims to combine the interpretability of the decision tree with the superior prediction power of the decision forest [31].
For this purpose, a decision forest is trained on a data set. As mentioned, the pruning function described in section 2.4.1 prunes this forest. A conjunction set is then derived, as explained in section 2.4.2. The conjunction set is then analyzed and built into a new decision tree using the following algorithm [31]:
The algorithm selects one rule R contained in the conjunction set and examines if it is possible to split the conjunction set on this rule.

The rule R that dictates how the conjunctions are split will be selected based on the most information gained. The information gained is calculated as follows:

$$IG(CS_i, R) = \frac{|CS_{i1}|\cdot\text{entropy}(CS_{i1})+|CS_{i2}|\cdot\text{entropy}(CS_{i2})}{|CS_{i1}|+|CS_{i2}|} - \text{entropy}(CS_i).$$

Here, CS_i is the conjunction set that will be split on the rule R. CS_{i1}, CS_{i2} are the newly created conjunction sets by dividing the initial conjunction set.
Entropy in probability theory and information theory quantifies the average level of uncertainty or information content inherent in a random variable's possible outcomes[36, 28]. For a discrete random variable X with possible outcomes $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p(x_1), p(x_2), \ldots, p(x_n)$, the entropy$(X)$ is defined as:

$$\text{entropy}(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i).$$

This expression captures the spread or unpredictability in the distribution [36, 28]. The more uniformly distributed the probabilities are, the higher the entropy. If all probability is accumulated in a

single outcome (maximum predictability), entropy is zero; if all outcomes are equally likely (maximum uncertainty), entropy is maximized [36, 28].
Entropy is a way to measure the disorder or randomness in a system and provides a way to assess the information to be gained from a system [36, 28].

If it is possible to split the conjunction set on R, it creates a node and gives it R and the threshold analogous to the description of the nodes of a decision tree [31, 23, 26].
The remaining conjunctions are divided by looking at the threshold [31]:

1. If a conjunction contains R and fulfills the threshold, it will be put on the true side of the node.
2. If it contains R but does not fulfill the threshold, it will be put on the false side of the node.
3. When a conjunction does not contain R, it will be put on both sides of the node.

If the remaining conjunctions all point to the same class or splitting the conjunction set would not reduce the number of conjunctions at the resulting sets, the node will be defined as a leaf. A prediction for new entries can now be made by passing through the tree from the root to a leaf and averaging the probabilities of the nodes [31].
This should lead to a model with a prediction power close to that of a decision forest and the interpretability of a decision tree [31].

2.6 Interpretable machine-learning models

In domains like health care, the use of machine-learning, as in all other fields, will increase. It demands interpretability so humans can decide if they want to trust the model decision and carry out the following step [21, 2].
Interpretable machine-learning models mean that the entire logic of these models can be understood by humans[13]. If the model can be visualized in a simple diagram or be described in plain text, interpretability is also archived [29].

2.6.1 Decision tree interpretability

The decision tree is a representative of an interpretable machine-learning model. It measures a threshold for a given characteristic of a data set [18].
For example, in the Fisher's Iris dataset, each row represents one of three related iris flower species. The decision tree classifies the three types of flowers by their length and the width of the sepals and petals [23].
If a decision tree is trained on this data, each inner node will contain a tested characteristic and a class distribution. Each leaf node contains just a vector with the class distribution:
First dimension: Iris setosa, second dimension: Iris virginica, and third dimension: Iris versicolor [23].
Figure 2.1 shows an example tree trained on this dataset. The non-leaf nodes have one attribute of the data, a threshold on which the decision will be made, the number of samples tested with this rule, and

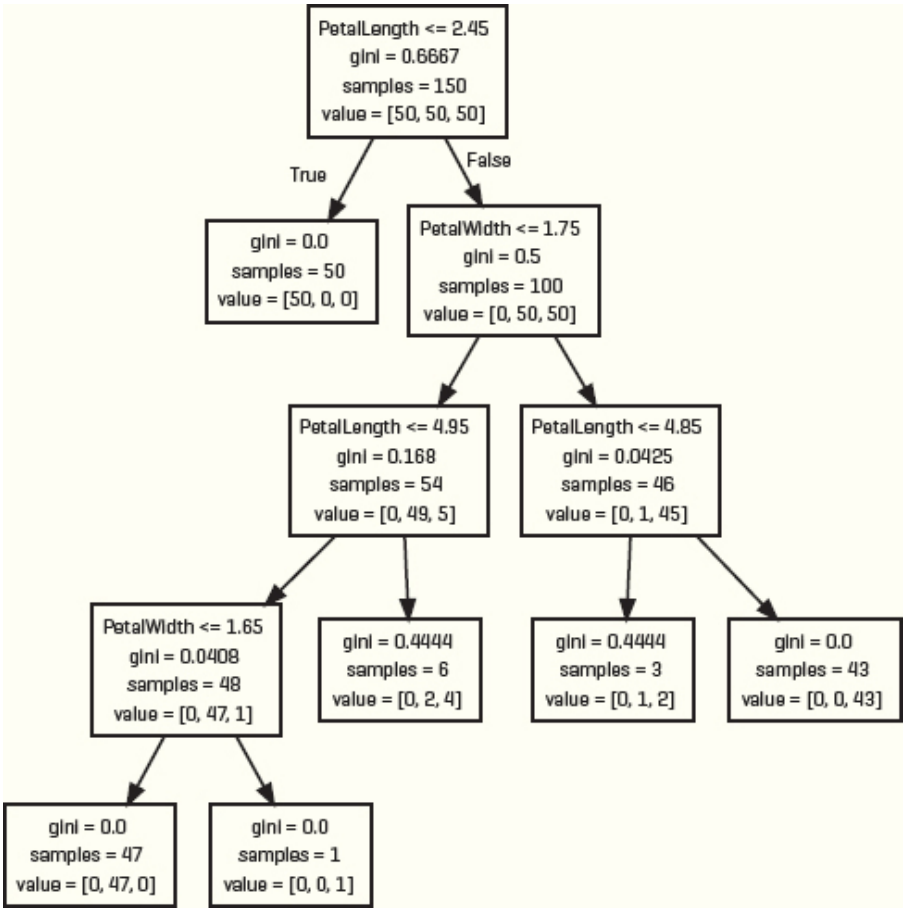


Figure 2.1: Visualisation of the decision tree classifier.
(Oana Niculaescu,2018) [23]

distribution into the classes [18].
Depending on their path from the root to the leaf, new entries are distributed into the classes. Since each leaf represents a clear set of characteristics, a classification can be easily understood [4, 25]. Due to this quality, decision trees are used in domains where understanding the decision process and its outcome is crucial. Additionally, decision trees can be used to explore domains to reveal new correlations that have not been discovered yet [7, 10].

2.7 Validation

2.7.1 Experiment execution

The experiment was implemented in Python, and a sci-kit-learn's balanced random forest classifier was used for the imbalanced set and the random forest classifier for the balanced set. The number of estimators that should be used for the forest was set to 100, and each of the trees in the forest was set to a maximum depth of five. Both versions had the maximum number of 3000 conjunctions per loop, and ten was set to be the minimum amount of estimators left after the pruning.
The model was trained on the above-described datasets that are based on the MIMIC-III dataset to classify the data into patients who have died and patients who have not died. It was also trained on the rate of postoperative delirium data where the model should classify the data into acquired delirium or no acquired delirium. Then, the forest was pruned, and conjunction sets were derived from them and assembled into the forest-based trees.
A sci-kit-learn decision tree that had the same parameter as one of the estimators of the random forest classifier was trained on the same data with the same task to compare whether it is sensible to spend the extra computing time to build a forest-based tree instead of the basic decision tree.
Using a sci-kit-learn stratified K-fold, the dataset was randomly split into five folds, where four folds were used to train. After every training run, the Area under the curve score of the receiver operating characteristic curve, the accuracy, and the predictions on the remaining test fold were measured. This was repeated 6 times for 30 runs of training and testing of the models.
In addition, after every training session, the forest-based and basic decision trees and their depth were recorded in a text file. This process was conducted for all of the three data sets.

2.7.2 Receiver operating characteristic curve

The receiver operating characteristic curve (ROC curve) is a plot of the true positive rate (TPR or sensitivity) on the y-axis and the false positive rate (FPR or $1-$ specificity) on the x-axis of a classification method.
Figure 2.2 is an example of an ROC curve. It is used to evaluate the predictive performance of a binary classification method. An ROC curve represents values of the average amount of sensitivity of the tested classification method for all specificity [22, 32].

The specificity and sensitivity of a flawless classification method would both be equal to one. If this method is displayed as an ROC curve, the curve would start at (0,0), immediately jump to (0,1), and then be a horizontal line to (1,1). A classification method aims to have an ROC curve near this example [6].
A classification model that would return an equal amount of false and true positives would have no predictive capability. Its ROC curve would be a straight line from (0,0) to (1,1). If the ROC curve is above this line, the classification model has a predictive capability [6].

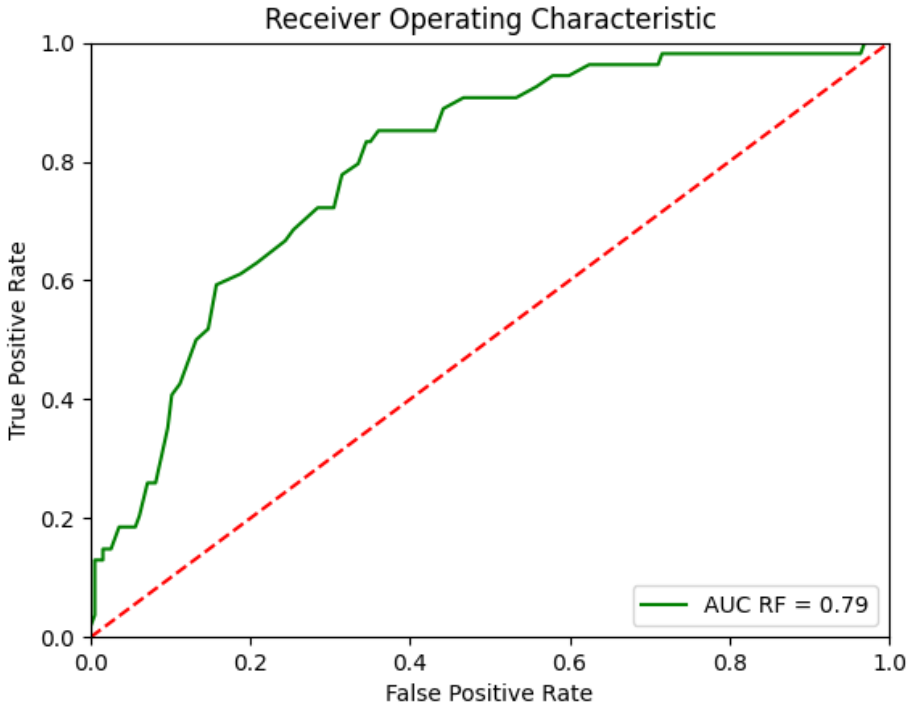


Figure 2.2: Example Receiver Operating Characteristic Curve.

2.7.3 Area under the curve

The area under the curve (AUC) is a measure to summarise the predictive capability of the classification model. An AUC of 0.5 means that the model is as good as guessing. If the AUC is above that, it means that the model has predictive capabilities[22, 32, 6].
The ROC curve of the classifier example depicted in figure 2.2 has an AUC value of 0.79 and would be a fair classification method by this interpretation.
The ROC curve is an effective metric for comparing the predictive ability of two or more classifier methods. The greater the AUC, the more suitable the classifier may be considered to classify the data correctly generally [22, 6].

2.8 MIMIC-III Dataset

MIMIC-III ('Medical Information Mart for Intensive Care') is a vast single-center database comprising information from a grand tertiary care hospital regarding patients treated in critical care units. Information of the vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more are contained in the data [17].
In this work, the admission table is used to list all the registered patients within a Hospital Admission, referred to as HAMID. The admission table contains a column that records whether the patient has died. The column's diagnosis, which may contain a diagnosis made in this hospital admission, and

Admission Type, which includes the type of admission, were also used. The HAMID then is used to join the admission table with the lab event table that contains lab results that have been recorded for the patients [17].
The admission table contains 58977 patients from which 5854 have died. To create a balanced dataset for the experiment to work on, the 5854 rows in which the patient has passed have been taken. Then, the remaining rows were randomly undersampled until a 1:1 ratio of died and not died had been reached.
To train the models on the complete data set as the imbalanced variant would skew the results because the models would have more information than the balanced data. To combat this, the died/not died ratio of the original data set is taken; both the ”died”-class and the ”not died”-class are randomly undersampled until the resulting set is the same size as the previously described balanced data set.

2.9 Rate of postoperative delirium dataset

The second dataset used for learning is also a medical data set. The rate of delirium after a long surgery and how the depth of anesthesia influences it were recorded in it [27].
Between March 2009 and May 2010, patients over the age of 60 who were scheduled to have at least 60 minutes of general anesthesia during their procedure are contained in this data [27].
A consecutive sample of 1277 patients was randomly split into two groups. For 638 of the subjects, authorization was granted to the anaesthesiologists to guide the anesthesia by utilizing the bispectral index (BIS) data. For the other 639, the authorization was not granted. From these subject’s data, 1155 were examined, of which 575 were guided by the BIS data, and 580 were not [27].
Their cognitive function was measured immediately after the surgical procedure, after one week and after three months [27]. In addition, the prescribed medication the patients were taking was also recorded, including the glucose blood levels and whether the patient was delirious or not [27]. 18.8% of the participating patients in the study were diagnosed with postoperative delirium, which means that the class distribution of this dataset is imbalanced with a similar but not as steep class ratio as the MIMIC-III dataset [27, 17].

3 Results

3.1 Prediction performance results of the three models

Figures 3.1a, 3.1c, 3.1b, 3.2a 3.2c, 3.2b, 3.3a, 3.3c and 3.3b all show the mean and standard deviation (SD) of the ROC curve, as well as the AUC with SD, for 30 runs of one of the three models on one of the three datasets. Each figure represents a combination of one model and one dataset, making up for the nine resulting figures. On each of them, all five folds were used successively for testing, and the other four folds in the run were used for training. This was repeated six times for a total of 30 runs.
The ROC curves of each figure rises above the red dashed diagonal line of random guessing. This indicates, that each model is performing better than random on every dataset, as they all achieve higher TPR at various levels of the FPR. The shaded area around the ROC curves represents the SD in TPR and FPR to visualize the range of model performance across different thresholds.

Figure 3.1a shows the described results for the forest. When run on the MIMIC-III derived imbalanced dataset, it achieves higher TPR at various levels of the FPR. It also archived higher TPR at various levels of the FPR then the forest-based tree and the basic tree on the same data. The shaded area indicates that performance is fairly stable, as the band is not very wide and smaller than the ones of the forest-based tree and the basic tree for the same data.
The mean AUC with SD for the forest is 0.81 ± 0.02.

Figure 3.1c shows the described results for the forest-based tree. When run on the MIMIC-III derived imbalanced dataset, it archived lower TPR at various levels of the FPR then the forest and the basic tree on the same data. The shaded area indicates that performance is stable, as the band is not wide and smaller than the one of the basic tree for the same data.
The mean AUC for the new model is 0.76 ± 0.04.

Figure 3.1b shows the described results for the basic tree. When run on the MIMIC-III derived imbalanced dataset, it achieves higher TPR at various levels of the FPR and higher than the forest-based tree on the same data. The shaded area indicates that performance is not very stable, as the band is wider and larger than the ones of the forest-based tree and the forest for the same data.
The mean AUC for the decision tree is 0.79 ± 0.02.

Figure 3.2a shows the described results for the forest. When run on the MIMIC-III derived balanced dataset, it archived higher TPR at various levels of the FPR than the forest on the MIMIC-III imbalanced data. The shaded area indicates that performance is fairly stable, as the band is not very wide and smaller than the ones of the forest-based tree and the basic tree for the same data. It is also smaller than the band of the forest on the MIMIC-III imbalanced data.
The mean AUC with SD for the forest is 0.84 ± 0.04.

Figure 3.2c shows the described results for the forest-based tree. When run on the MIMIC-III derived balanced dataset, it archived lower TPR at various levels of the FPR than the forest and the basic tree on the same data and about the same as the forest-based tree on the MIMIC-III Imbalanced data. The shaded area indicates that performance is stable, as the band is not wide. It is larger then the ones of the forest and the basic tree for the same data. It is about the same as the band of the forest-based tree on the MIMIC-III imbalanced data.
The mean AUC for the new model is 0.76 ± 0.04.

Figure 3.2b shows the described results for the basic tree. When run on the MIMIC-III derived balanced dataset, it also archived higher TPR at various levels of the FPR than the forest-based tree on the same data and higher than the basic tree on the MIMIC-III Imbalanced data. The shaded area indicates that performance is fairly stable, as the band is not very wide and smaller than the one of the forest-based tree for the same data. It is also much smaller than the band of the basic tree on the MIMIC-III imbalanced data.
The mean AUC for the decision tree is 0.83 ± 0.05.

Figure 3.3a shows the described results for the forest. When run on the rate of postoperative delirium imbalanced data, it archived higher TPR at various levels of the FPR than the forest-based tree and the basic tree on the same data and lower than the forests on both of the MIMIC-III datasets. The shaded area indicates that performance is fairly stable, as the band is not very wide and smaller than the ones of the forest-based tree and the basic tree for the same data. It is larger than the band of the forest on both of the MIMIC-III datasets.
The mean AUC with SD for the forest is 0.74 ± 0.05.

Figure 3.3c shows the described results for the forest-based tree. When run on the rate of postoperative delirium imbalanced data, it archived higher TPR at various levels of the FPR than the basic tree on the same data and lower than the forest-based tree on both of the MIMIC-III datasets. The shaded area indicates that performance is stable, as the band is not wide and smaller than the ones of the basic tree for the same data. It is about the same as the band of the forest-based tree on both of the MIMIC-III datasets.
The mean AUC for the new model is 0.70 ± 0.06.

Figure 3.3b shows the described results for the basic tree. When run on the rate of postoperative delirium imbalanced data, it archived lower TPR at various levels of the FPR than the forest and the forest-based tree on the same data lower than the basic tree on both of the MIMIC-III datasets. The shaded area indicates that performance is not very stable, as the band is wide and larger than the ones of the forest and the forest-based tree for the same data. It is larger than the band of the basic tree on the MIMIC-III balanced data and about the same on the MIMIC-III imbalanced data.
The mean AUC for the decision tree is 0.65 ± 0.05.

3.2 Features of basic and forest-based tree

The features of the 30 recorded forest-based and basic decision trees were extracted from the files, and their occurrence was counted.
Table 3.1 shows the ten features with the most appearances and the percentage of these appearances on the MIMIC-III imbalanced dataset for the forest-based and basic decision tree. The features of the trees do not overlap at all on the MIMIC-III imbalanced dataset.
The results in table 3.2 represent the forest-based and basic decision tree's ten most prevalent features and their percentage when trained on the MIMIC-III balanced data set. The features of the trees overlap in the feature "PT" on the MIMIC-III imbalanced dataset.
The ten features that were contained in the forest-based and basic decision trees most frequently, when trained on the rate of postoperative delirium imbalanced data and their percentages are recorded in table 3.3. The features of the trees overlap in the features "HB min", "Glucose min", "Age in years", "Hb max" on the rate of postoperative delirium imbalanced data.
In table 3.4, the total amount of features that impacted the decision-making process of the forest-based and basic decision trees are shown and differentiated between the MIMIC-III datasets and the rate of postoperative delirium imbalanced data.
Table 3.4 shows that the forest-based tree used 117 while the basic decision tree used 62 features on the MIMIC-III imbalanced dataset. The forest-based tree used 180 while the basic decision tree used 45 features on the MIMIC-III balanced dataset. The forest-based tree used 29 while the basic decision tree used 22 features on the rate of postoperative delirium imbalanced data.
The features that were in the top ten of the forest-based tree on both of the MIMIC-III datasets are "PT", "Potassium", "MCH", and "MCHC".
The features that were in the top ten of the basic tree on both of the MIMIC-III datasets are "Urea Nitrogen", "White Blood Cells", "Base Excess", "Temperature", "ADMISSION TYPE=EMERGENCY", "Platelet Count", and "Oxygen".
The features that were both in the top ten features of the forest-based tree and the basic tree on one data set are indicated by bold text. Similarly, features that were in the top ten of the forest-based tree on both of the MIMIC-III datasets are underlined, as are those of the basic tree.

3.3 Average depth

Table 3.5 depicts the mean depth of the models when trained on the MIMIC-III imbalanced dataset, the MIMIC-III balanced data set, and the imbalanced dataset pertaining to the incidence of postoperative delirium. On the MIMIC-III imbalanced dataset, the forest depth was 2.19, the new model depth was 12.86, and the basic tree depth was 5.0. On the MIMIC-III balanced dataset, the forest depth was 4.26, the new model depth was 11.42, and the basic tree depth was 5.0. On the rate of postoperative delirium imbalanced data, the forest depth was 3.41, the new model depth was 14.34, and the basic tree depth was 5.0.

New model feature	Percentage	Decision tree feature	Percentage
PT	7.50	Urea Nitrogen	14.17
Potassium	5.27	White Blood Cells	12.36
MCH	4.24	Base Excess	7.53
Red Blood Cells	3.94	Temperature	6.73
Basophils	3.78	ADMISSION TYPE=EMERGENCY	6.33
Eosinophils	3.23	Specific Gravity	5.92
MCV	3.19	Platelet Count	4.92
Creatinine	3.14	PEEP	4.32
MCHC	2.97	Required O2	3.31
Chloride	2.83	Oxygen	2.71

Table 3.1: Most prevalent features of the trees on the imbalanced MIMIC-III data. Bold: Features that were both in the forest-based tree and the basic tree. Underlined: Features that were in both of the models when trained on the MIMIC-III data.

New model feature	Percentage	Decision tree feature	Percentage
MCH	4.94	White Blood Cells	16.23
PT	**4.90**	Urea Nitrogen	11.73
Oxygen	4.64	Base Excess	7.53
MCHC	3.92	Osmolality, Urine	6.44
Free Calcium	3.43	ADMISSION TYPE=EMERGENCY	5.98
Potassium	3.38	Bicarbonate	5.82
pO2	3.25	Temperature	4.89
Specific Gravity	3.00	**PT**	**4.42**
Glucose	2.95	Platelet Count	3.96
Potassium, Whole Blood	2.93	Oxygen	3.18

Table 3.2: Most prevalent features of the trees on the balanced MIMIC-III data. Bold: Features that were both in the forest-based tree and the basic tree. Underlined: Features that were in both of the models when trained on the MIMIC-III data.

New model feature	Percentage	Decision tree feature	Percentage
Geschlecht	9.71	**Age in years**	**17.20**
Hb min	**8.24**	**Hb min**	**15.00**
ASA	6.35	**Glucose min**	**13.30**
aufDesflurane	5.65	Glucose max	11.44
betaBlocker	5.59	**Hb max**	**10.42**
Glucose min	**5.55**	einlThiopental	5.50
Age in years	**5.35**	einlPropofol	5.33
aceAntagonist	5.35	ASA	2.62
Hb max	**5.30**	antidepressiva	2.28
antiPlatelet	5.11	betaBlocker	2.11

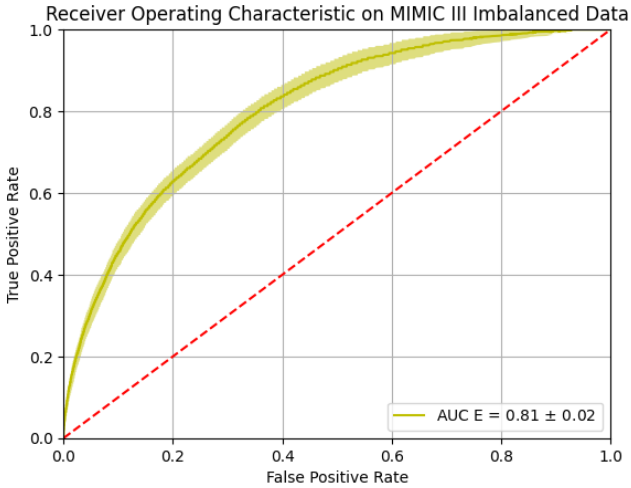
Table 3.3: Most prevalent features of the trees on the rate of postoperative delirium imbalanced data. Bold: Features that were both in the forest-based tree and the basic tree.

Dataset	New model features	Decision tree features
MIMIC-III Imbalanced	117	62
MIMIC-III balanced	180	45
Rate of delirium	29	22

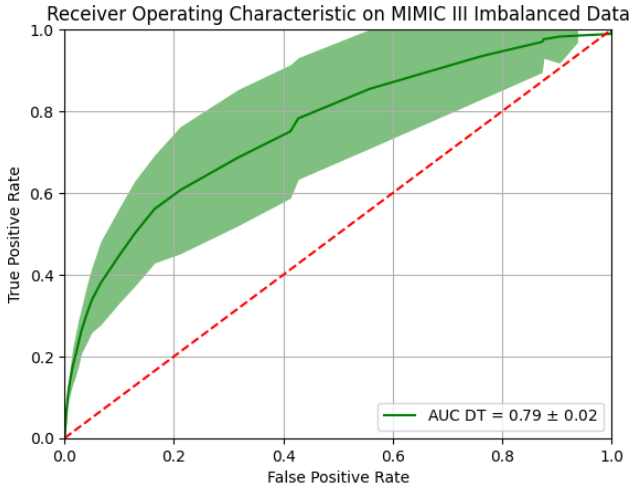
Table 3.4: Total amount of features in the trees.

Data set	Ensemble average depth	New model average depth	Decision tree average depth
MIMIC-III imbalanced	2.199015	12.864285	5.0
MIMIC-III balanced	4.267578	11.422735	5.0
postoperative delirium imbalanced	3.416793	14.348827	5.0

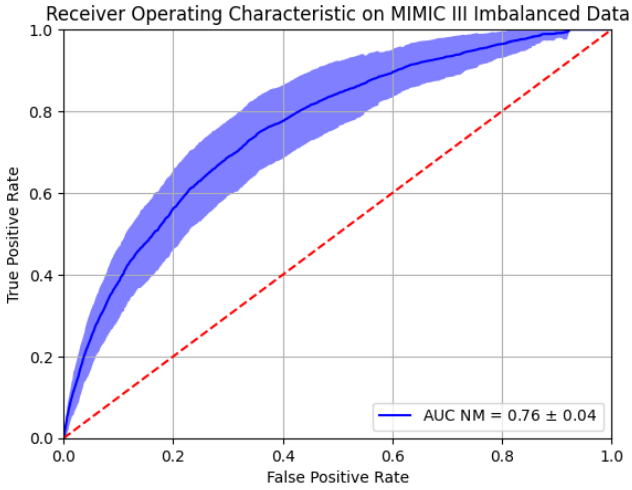
Table 3.5: Average depth of the models on the different datasets.



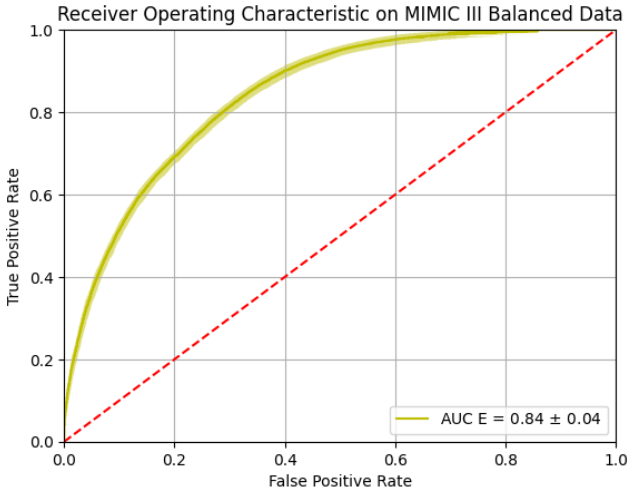
(a) Mean with standard deviation of ROC and AUC for 30 runs of the ensemble forest when trained on four folds of MIMIC III imbalanced data and tested on one fold.



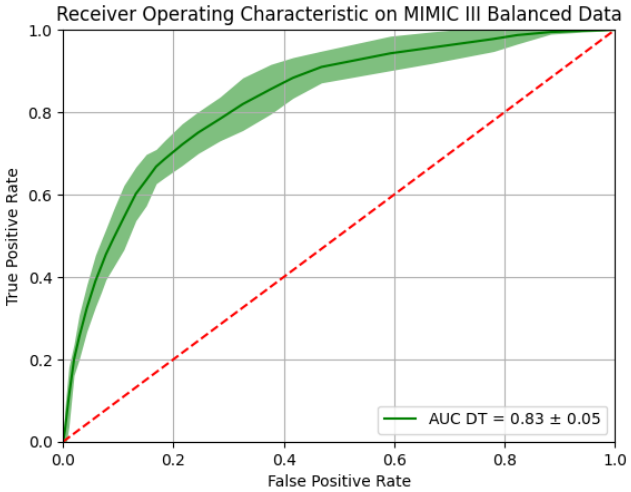
(b) Mean with standard deviation of ROC and AUC for 30 runs of the decision tree when trained on four folds of MIMIC III imbalanced data and tested on one fold.



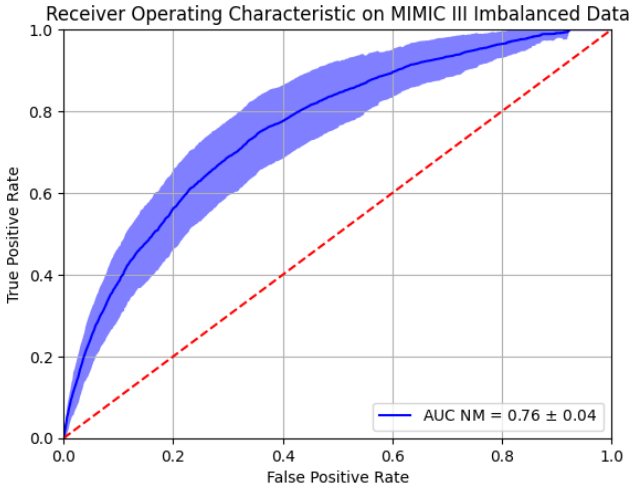
(c) Mean with standard deviation of ROC and AUC for 30 runs of the new model when trained on four folds of MIMIC III imbalanced data and tested on one fold.



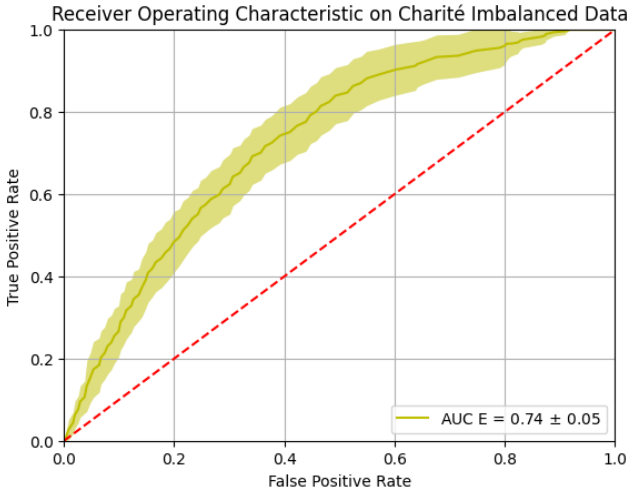
(a) Mean with standard deviation of ROC and AUC for 30 runs of the ensemble forest when trained on four folds of MIMIC III balanced data and tested on one fold.



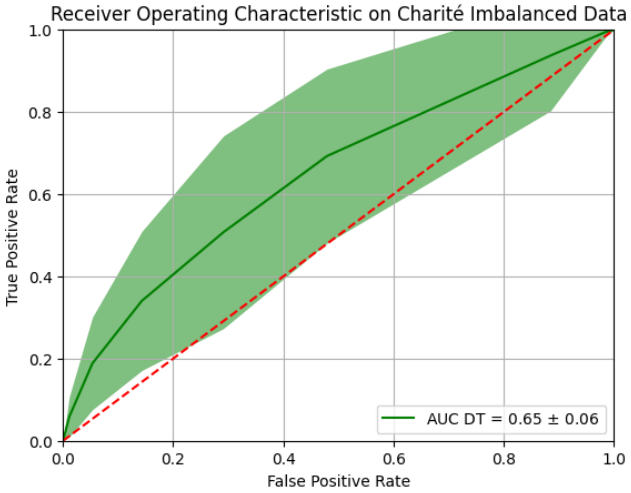
(b) Mean with standard deviation of ROC and AUC for 30 runs of the decision tree when trained on four folds of MIMIC III balanced data and tested on one fold.



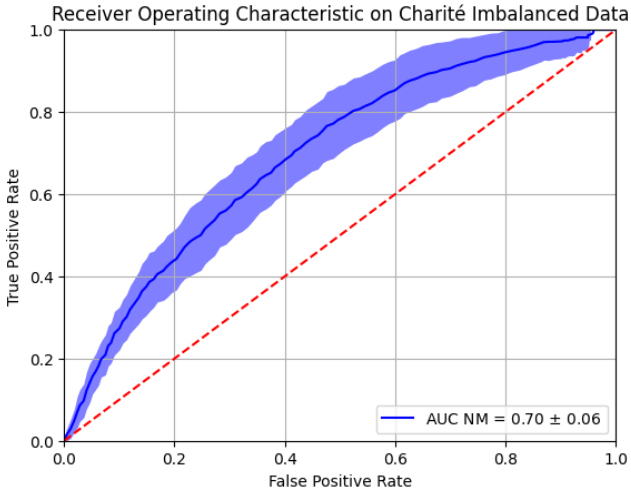
(c) Mean with standard deviation of ROC and AUC for 30 runs of the new model when trained on four folds of MIMIC III balanced data and tested on one fold.



(a) Mean with standard deviation of ROC and AUC for 30 runs of the ensemble forest when trained on four folds of the rate of postoperative delirium imbalanced data and tested on one fold.



(b) Mean with standard deviation of ROC and AUC for 30 runs of the decision tree when trained on four folds of the rate of postoperative delirium imbalanced data and tested on one fold.



(c) Mean with standard deviation of ROC and AUC for 30 runs of the New Model when trained on four folds of the rate of postoperative delirium imbalanced data and tested on one fold.

4 Discussion of results

The results printed in the figures 3.1a, 3.1c and 3.1b suggest that the forest-based tree will get outperformed by the two other models when they are trained on the MIMIC-III derived datasets. The postoperative delirium imbalanced data results imply that the forest-based tree 3.3c can perform better than a basic tree 3.3b. However, the results of the three models on the MIMIC-III derived balanced dataset presented in figures 3.2a, 3.2c and 3.2b, also support the claim that using a basic decision tree is more sensible not just in terms of computation time and resource usage but also in prediction power.

In general, the forest-based tree looks at more features than the basic decision tree, which is consistent with the resulting sizes of the trees. In table 3.4 the number of features included in the forest-based tree are compared to those present in the basic decision tree. The larger the discrepancy in numbers is the poorer is the performance of the forest-based tree in terms of prediction power. Conversely, when the number of features is similar in both trees, the forest-based tree demonstrates a better performance, as evidenced in the postoperative delirium imbalanced data row of table 3.4.

When looking at the features that impacted the decision-making process the most on the MIMIC-III data, it seems that the forest-based tree and the basic decision tree recognize different features as important. This is the most noticeable on the MIMIC-III Imbalanced data in table 3.1, where the ten most often used features to make a decision do not overlap. This is also the case on the MIMIC-III balanced data in table 3.2; here, the two models have just one common feature. This may be a good way to discover why the forest-based tree performed poorly on the MIMIC-III data.
On the rate of postoperative delirium imbalanced data, the two models still had differences in the feature's impactfulness in table 3.3, but this time, they had more overlap. On this dataset, the two models had four features in common. The forest-based tree also performed better on this dataset, which suggests that the selection of essential features worked better this time.
It is also interesting to see that the forest-based tree tends to distribute the features more evenly than the basic tree and that the dataset where the forest-based tree chooses features to be more prominently represented is performing better.

The forest-based trees are also much more complex than the basic decision trees, as the results in table 3.5 suggest. This can be explained by the way they are created. They contain information about ten decision trees; therefore, they need to be more significant. However, there are still at least six decisions deeper than the basic tree, meaning the forest-based trees are 64 times larger than the basic decision tree if they are entirely filled. This means that they need more space to store and take longer to produce results than a bare tree, which supports the usage of the basic tree in terms of explainability.
All the results show that the forest-based tree has less deviation than the basic tree in terms of sensitivity based on imbalanced data seen in figures 3.1c, 3.3c, 3.1b, 3.3b. This is because a decision tree looks at the data not as a whole and tries to find a local maximum, whereas the forest-based tree is de-

rived from a forest that can find global maxima in the data that then is represented in the forest-based tree, which leads to more reliability [19].

According to Omer Sagis and Lior Rokach's paper [31], the usage of a forest-based tree is most effective when there is a large gap between the performance of the random forest and a basic tree because then the forest-based tree has a lot of space to be more accurate than the basic tree but still be more explainable than the forest [31].
The results of this work disagree with this claim. As the gap between the forest depicted in figure 3.2a and the basic decision tree depicted in figure 3.2b on the MIMIC-III balanced data is about as much as the gaps shown in their work yet the forest-based tree depicted in figure 3.2c is performing worse than the basic tree.
On the imbalanced data, the gap between the forest depicted in figure 3.1a and the bare decision tree depicted in figure 3.1b is more significant, and still, the forest-based tree depicted in figure 3.1c performs worse than the basic tree. This contrasts with the original, which claims that the best practice would be a forest-based rest-based tree for classification tasks where the decisions are interpretable.
On the postoperative delirium imbalanced data, the gap between the forest shown in figure 3.3a and the basic tree shown in figure 3.3b is significantly larger than on both of The MIMIC-III datasets. Here, the forest-based tree shown in figure 3.3c performs as the authors expected. Despite that, the forest-based tree seems far less effective than the original paper claims [31] due to the fact that it performs worse than the basic tree on two of the test datasets.

Despite the promising claims of the forest-based tree, it performed poorly on both MIMIC-III datasets. Surprisingly, it performed rather well with the imbalanced postoperative delirium data.
The difference was unexpected because the MIMIC-III datasets were almost 10 times larger than the postoperative delirium imbalanced dataset in patients alone. The MIMIC-III data also had many more events performed on the patients than postoperative delirium imbalanced data. The forest-based tree was not able to find the relevant date in the MIMIC-III data. The postoperative delirium imbalanced data was more focused for the given classification task at hand and because of that the forest-based tree was able to perform better [27].
The different outcomes imply that the performance of the forest-based tree depends more on the structure of the data than on the sheer amount.

It is interesting to note that the forest-based tree has significantly less deviation in the imbalanced data than the basic decision tree. However, the forest-based tree has the same deviation as the imbalanced one in the balanced data. In contrast, the basic tree has much less deviation on the balanced data in terms of the ROC curve than on the imbalanced data. It also seems that the amount of deviation that the forest has does not impact the forest-based tree's deviation.
The forest on the MIMIC-III balanced data has the least deviation. Still, the derived forest-based tree has a similar deviation as the forest-based trees derived from the forests trained on the MIMIC-III imbalance data and the postoperative delirium imbalanced data.

In examining the most prevalent characteristics of forest-based and basic decision trees, a notable finding emerged. The most significant features of the basic decision tree exhibited considerable consistency across both imbalanced and balanced MIMIC-III datasets, with seven of the ten most prevalent features being shared. In contrast, the most prevalent features of the forest-based tree only

share four common features.
This indicates that the process of determining which feature to include in the tree is less consistent than that in the basic decision tree, resulting in a less accurate model.

4.1 Limitations

The results demonstrated that the predictive capability of the forest-based tree was insufficient when evaluated using the MIMIC-III dataset. However, the tree exhibited the anticipated performance when assessed on the postoperative delirium imbalanced dataset. In light of the conflicting results, it is not possible to make a general statement regarding the expected performance of the forest-based tree in comparison to a basic decision tree. Furthermore, a reliable statement regarding the feature selection of the forest-based tree cannot be made, as the experiment was not designed to capture this aspect of the forest-based tree.

5 Conclusion and future work

This study examined the findings of the research conducted by Omer Sagis and Lior Rokachs [31] and sought to determine whether forest-based trees were suitable for use in imbalanced medical data to develop an interpretable classification method with greater predictive power than a basic decision tree.

A systematic comparison of the performance of random forests, forest-based trees, and basic decision trees was conducted on two medical data sets that were imbalanced and balanced, respectively, and derived from the same MIMIC-III dataset. Additionally, a third data set containing postoperative delirium data, which was also imbalanced, was included in the analysis.
The forest-based tree exhibited inferior performance compared to the basic decision tree with regard to prediction power and interpretability on the two MIMIC-III datasets. However, it demonstrated superior performance in terms of prediction power on the postoperative delirium data. Consequently, a general argument regarding the usage of the forest-based tree method in the background of imbalanced medical data cannot be made.

The findings of this study indicate that the forest-based tree method is less effective than previously anticipated. Additionally, the predictive performance of the forest-based tree is found to be dependent upon the data used for training. Furthermore, the study reveals that the process of selecting the features on which the forest-based tree makes its decisions is less consistent than the process applied by the basic decision tree. The decision-making process of the forest-based tree is less consistent than that of the basic decision tree. Furthermore, the difference in the number of features considered during the decision-making process between the two tree types seems to be influencing the performance of the forest-based tree.

Despite the fact that this thesis examined the usage of forest-based trees based on medically imbalanced data, it is acknowledged that there are limitations to this research that could be further investigated in the future. First, an investigation could be conducted into why the divination of the ROC curve of the forest-based tree appears to be independent of the deviation of the forest ROC curve from which it was derived. Secondly, an investigation into the data dependence of the forest-based tree could be conducted to discover which data structure enhances the predictive performance of the forest-based tree. Finally, an examination of the relationship between the influence of the number of features integrated into the trees and the forest-based tree's performance and the apparent inconsistency in the process that determines the features to be included in the forest-based tree could be undertaken.

In conclusion, the forest-based tree is not generally the optimal choice for classifying imbalanced medical data in an interpretable manner. A basic decision tree can prove to be more effective and require less computation time and storage space. An examination on the data is necessary to determine the advantages of using the forest-based tree.

DeepL

DeepL was used for correction in this work. In doing so, the AI tool adapted the grammar and spelling, but the content remained unchanged.
URL: https://www.deepl.com

Grammarly

Grammarly was used for correction in this work. In doing so, the AI tool adapted the grammar and spelling, but the content remained unchanged.
URL: https://app.grammarly.com

Bibliography

[1] *Avoiding Overfitting of Decision Trees*, pages 119–134. Springer London, London, 2007.

[2] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, page 559–560, New York, NY, USA, 2018. Association for Computing Machinery.

[3] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

[4] Chidanand Apté and Sholom Weiss. Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 13(2):197–210, 1997. Data Mining.

[5] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, 2008.

[6] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 13: receiver operating characteristic curves. *Critical care*, 8:1–5, 2004.

[7] I. Bratko. Machine learning: Between accuracy and interpretability. In Giacomo Della Riccia, Hans-Joachim Lenz, and Rudolf Kruse, editors, *Learning, Networks and Statistics*, pages 163–177, Vienna, 1997. Springer Vienna.

[8] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 18, New York, NY, USA, 2004. Association for Computing Machinery.

[9] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, jun 2004.

[10] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8, 1995.

[11] Chris Drummond and Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 01 2003.

[12] Alex A. Freitas. Comprehensible classification models: a position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10, mar 2014.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[14] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263 – 1284, 2009. Cited by: 6730.

[15] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[16] Qinghua Hu, Daren Yu, Zongxia Xie, and Xiaodong Li. Eros: Ensemble rough subspaces. *Pattern Recognition*, 40(12):3728–3739, 2007.

[17] Alistair EW 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. *Scientific data*, 3(1):1–9, 2016.

[18] Sotiris B Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39:261–283, 2013.

[19] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, nov 2016.

[20] Charles X. Ling and Chenghui Li. Data mining for direct marketing: problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 73–79. AAAI Press, 1998.

[21] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[22] Jayawant N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[23] Oana Niculaescu. Classifying data with decision trees. *XRDS*, 24(4):55–57, July 2018.

[24] Ioannis Partalas, Grigorios Tsoumakas, and Ioannis Vlahavas. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *ECAI 2008*, pages 117–121. IOS Press, 2008.

[25] J Ross Quinlan. Generating production rules from decision trees. In *ijcai*, volume 87, pages 304–307. Citeseer, 1987.

[26] J. Ross Quinlan. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72, 1996.

[27] F.M. Radtke, M. Franck, J. Lendner, S. Krüger, K.D. Wernecke, and C.D. Spies. Monitoring depth of anaesthesia in a randomized trial decreases the rate of postoperative delirium but not postoperative cognitive dysfunction. *British Journal of Anaesthesia*, 110:98–105, 2013.

[28] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. `why should i trust you?`: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[30] Lior Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016.

[31] Omer Sagi and Lior Rokach. Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61:124–138, 2020.

[32] Nahm Francis Sahngun. Receiver operating characteristic curve: overview and practical use for clinicians. *kja*, 75(1):25–36, 2022.

[33] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687 – 719, 2009. Cited by: 1253.

[34] Grigorios Tsoumakas, Ioannis Partalas, and Ioannis Vlahavas. A taxonomy and short review of ensemble selection. In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications*, pages 1–6, 2008.

[35] Anneleen Van Assche and Hendrik Blockeel. Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 418–429, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[36] Alfred Wehrl. General properties of entropy. *Rev. Mod. Phys.*, 50:221–260, Apr 1978.

[37] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, feb 2019.

[38] Yi Zhang, Samuel Burer, W Nick Street, Kristin P Bennett, and Emilio Parrado-Hernández. Ensemble pruning via semi-definite programming. *Journal of machine learning research*, 7(7), 2006.