Towards an Intelligent Preoperative Surgical Decision Support System: A machine learning-based approach

Cheima Bouden[a], Chaker Mezioud[b]

[a]*LISIA Laboratory, Abdelhamid Mehri University – Constantine 2, Algeria*

New City - Ali Mendjeli. Constantine, Algeria

cheima.bouden@univ-constantine2.dz

[b]*LISIA Laboratory, Abdelhamid Mehri University – Constantine 2, Algeria*

New City - Ali Mendjeli. Constantine, Algeria

chaker.mezioud@univ-constantine2.dz

***Abstract*— Artificial Intelligence has experienced a new impetus since the beginning of this decade under reasons interpreted by the increase in computing capacities, the emergence of distributed massive data processing methods, and powerful machine learning algorithms. Different sectors are showing signs of investing in this new technology, particularly the surgical field. In this area, the managers have found that the operating room is one of the most decisive resources in a hospital establishment, and that the automation and optimization of its process is a primary priority, in particular the preoperative phase, a phase deterministic in relation to the importance of patient care, as well as in relation to good management of the operating room. Through this article, the idea leads us to embark on the trail of proposing an intelligent decision support system for the surgical preoperative phase, based on Machine Learning models, for the reason of the ability to its algorithms to make precise and interpretable classifications, our choice fell on the "Random Forest" algorithm, while exploiting preoperative predictive data, namely in current surgery (blood pressure, oxygen pressure, cerebral activity, body temperature, blood sugar, hematocrit) or to come (detection of circulating tumor cells, .. etc). The proposed model will be validated through a Dataset approved by a renowned institutional review board, in the Python environment version 3.9.16 with Google Colab.**

***Keywords* — Artificial intelligence, Surgery, Preoperative phase, Machine Learning, Random Forest.**

I. INTRODUCTION

The operating room is the fundamental axis of a hospital, both is an intersection of a wealth of different material resources, and human with their distinct skills. So optimizing its operation is one of the primary concerns of hospital managers. This trend towards the pooling of resources puts even more emphasis on the need for patient care in order to reduce their risk, which is why it is necessary to take care of the preoperative phase of the surgical process, which is supposed to be the decisive step in the operating process. The objective of this paper is to move towards an intelligent decision support system for the preoperative phase before moving the patient to the operating room, an idea with a double interest from a medical point of view: reducing the patient's risk, while moving towards a dynamic and optimized planning of the operating room. The proposed approach will be based on Machine Learning models, and consists in developing algorithms capable of learning automatically from data and improving their performance. On this, we will rely on the "Random Forest" algorithm [7], a very efficient algorithm for handling high-dimensional data sets with complex nonlinear relationships between variables [15]. It should be noted that each surgical intervention is planned individually while using preoperative predictive data to provide the most suitable planning of the activity of the operating room. The proposed system must therefore first have integrated all the parameters that come into play; it will also have to be trained, via supervised machine learning with Datasets of real incident reports [10], in order to recognize risky situations. This paper is composed of different sections. Our study will be initiated by a state of the art in the surgical field, starting to examine the context of the operating room in hospitals, then its interaction with the various departments, and we are particularly interested in the preoperative phase. In another section we will unveil the field of AI (Artificial Intelligence), through its main derivative which is Machine Learning (Automatic Learning), while emphasizing Radom Forest algorithms. Our approach will be presented in a separate section, which will be illustrated by a case study. At the end, a general conclusion and outlook will also be discussed.

II. THE OPERATING ROOM : CONTEXT AND CHALLENGES

The operating room is the dense core of a hospital in terms of number and type of resources. He constantly interacts with different medical sectors, services and activities.

The operating room is at the interface of many activities: surgery, obstetrics, anesthesia, functional explorations, radiology and biology. The following figure summarizes all the different existing interactions [4]:

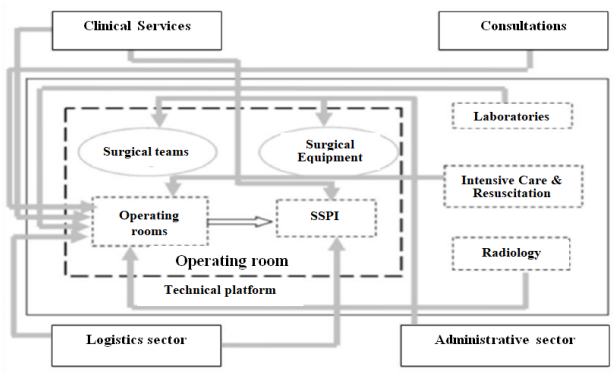

Fig. 1 Interaction of the Operating Room with other services

A. Main categories of surgery

In the operating room, there are three categories of surgeries: elective, ambulatory and urgent [4].

- Elective surgeries: concern operations planned following a surgical consultation. The patient is usually hospitalized in a care unit.
- Ambulatory surgeries: concern operations that do not require hospitalization.
- Urgent surgeries: these operations are unplanned and arrive at random in the operating room. They come either from a hospitalization unit, intensive care unit or from the emergency department.

A. Preoperative phase

The preoperative phase extends from the psychic preparation of the patient, before the intervention, until his exit from the recovery room. After a preoperative assessment and once favorable, the patient will then be anesthetized and finally operated by a surgical team. After his operation, he is transferred to the recovery room and remains there until the anesthesiologist authorizes his transfer to his room or, in the event of complications, he can be transferred to the intensive care and resuscitation unit [17]. We can illustrate this phase through the following figure:

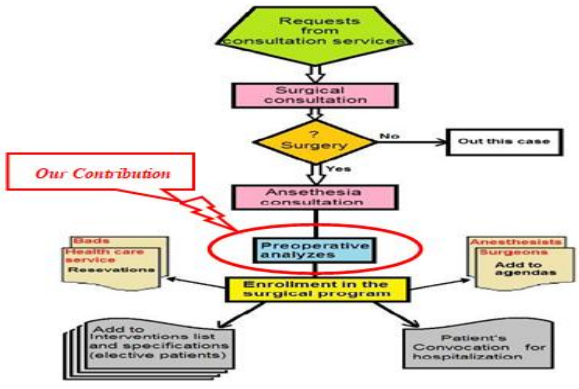

Fig. 2 Management of the preoperative phase

We have tried to target through the previous figure (fig 2) the circle of intervention of our proposed solution, the detailed approach with its different stages will be discussed later.

III. Machine Learning

Machine learning is the development of algorithms that can automatically learn from data and improve their performance over time. These algorithms can be used for a wide range of tasks, and its applications have established themselves in various industries, including finance, marketing, and healthcare [2].

A. Stages of Machine Learning

Machine learning involves a set of steps that are usually followed to create a machine learning model [16], which we can summarize them as follows:

- Data acquisition.
- Data pre-processing.
  - Data augmentation.
  - Data processing.
- Feature extraction.
- Division of data.
- Classification.
- Model evaluation using metrics.
- Model testing.
- Refinement of the model.

The figure below (Fig.3) summarizes the principle of automatic learning:

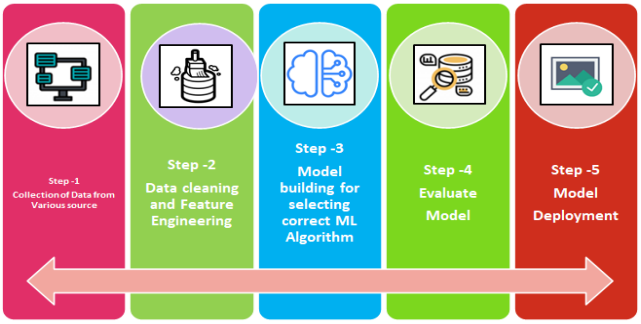

Fig. 3 Principle of machine learning [16]

B. Random Forest

Random Forest is a popular ensemble learning algorithm that combines multiple decision trees to create a more robust and accurate model. The first step in the Random Forest algorithm is the random selection of variables or features that will be used to construct each decision tree. This selection is made for each tree in the forest independently of other trees, and it is made from a random subset of all available variables.

A second step consists in the construction of decision trees: There are many algorithms to build a decision tree, such as: ID3, C4.5, C5, and CART (Classification and Regression Trees) [15]. We propose to use the CART algorithm as the basic learning algorithm, it is a binary recursive partitioning algorithm, because binary decision trees are faster to train and evaluate compared to the tree multi-way decision making. The decision tree can be visualized as a series of "if-else" statements that divide the input space into regions, where each region corresponds to a leaf node with a predicted value. Then a step of random feature selection, this process is called feature engineering, which consists in selecting the most useful features on which the model trains [12]. Another step is building multiple decision trees; the idea behind random forest is to build a large number of decision trees, where each tree is built using a random subset of the training data and a random subset of features. At the end, a Majority Vote will be applied. The principle of classification in the context of the random forest is to predict the class label of a new instance based on the class labels predicted by each decision tree in the forest. The class label that receives the most votes from the decision trees is then taken as the final prediction. This is called the principle of majority voting.

We can illustrate the principle of Random Forest through the following figure (Fig. 4):

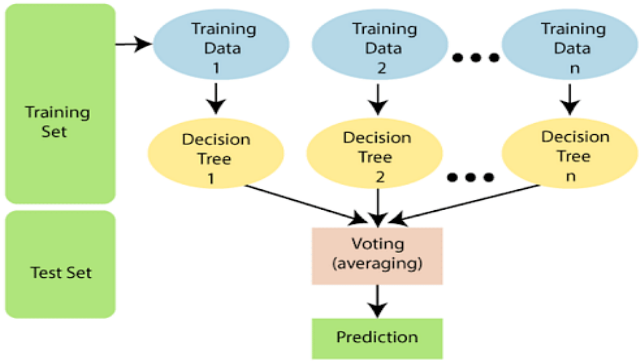

Fig. 4 Principal of Random Forest algorithm

C. *Machine Learning in the Medical Field*

The medical field is an area where machine learning is increasingly used. With the vast amount of medical data available, machine learning algorithms can be used to improve patient outcomes, diagnose diseases, and identify risk factors [20]. On this we will reserve the next section to present an approach based on machine learning, dedicated to the preoperative decision support of patients.

IV. Proposed Approach

Through this article we want to move towards an intelligent decision support system for the preoperative surgical phase, based on machine learning for a reliable and precise exploration and classification of preoperative patient assessments. We chose the "Random Forest" algorithm, while exploiting preoperative predictive data, namely in current surgery (blood pressure, oxygen pressure, brain activity, body temperature, blood sugar, and hematocrit) or going towards extension (detection of circulating tumor cells, etc.) [10]. The system can also help for better patient care, as well as intelligent and reliable operation of the operating room, through the perspective of moving towards dynamic and real-time planning. It should be noted that this future intelligent system would be requested by various hospitals and clinics, which will communicate their preoperative data. This system would be able to build a large Dataset, and after a training phase it will be able to guarantee an accurate prediction of the preoperative phase. Our primary concern is to ensure better management of the patient who will be operated on (risk reduction), which is why we have used the "Random Forest" model. This idea will ease the load for the different hospital units and unify the modality and credibility of this important preoperative process. It should also be noted that this system will contribute to a dynamic planning of the programming of the operating room, following any cancellation of a programming for a surgical intervention; this planning update will be communicated in real time.

A. *General Principle of the Approach*

The general idea of our future system is that through a predefined preoperative assessment (Blood assessment & questionnaire), which will be unified across all units and hospital structures (hospitals and clinics), which want to adopt this new system; we are going to move from the old culture "Assessment for All" to "Reflective Action". Because we have found that being satisfied with a systematic assessment can create a risk for the patient, as well as postoperative events. So through the acquisition of an important preoperative Dataset [10], we will train our model and which would be able through this phase, to predict the operative case of any new patient (New instance). The figure below globally illustrates the idea of our approach:

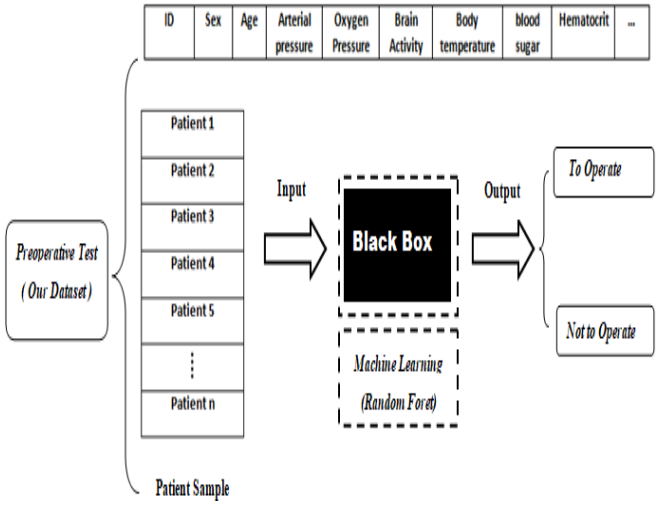

Fig. 5 General Description of the Approach

B. *Detailed Description of the Approach*

Once you have acquired the various information communicated from the various hospital units. And After going through the different main stages of machine learning; namely the preprocessing of the data and the extraction of the relevant characteristics that will guide towards an accurate prediction. We will subdivide our Dataset [10] into two essential parts: one reserved for training and the other reserved for testing. By using the principle of "Random Forest", we will subdivide the part reserved for training our Dataset into several sub-Datasets in a random manner and according to different partitions.

We will have several decision trees, according to different structures unlike the choice of the root attribute each time [7], [12]. So for any new instance to want to test, our model will be able to guarantee us several predictions according to the different routes traveled by the different decision trees, then a majority vote will be applied in order to decide the final prediction. The following figure (Fig. 6) will summarize the principle of our approach:

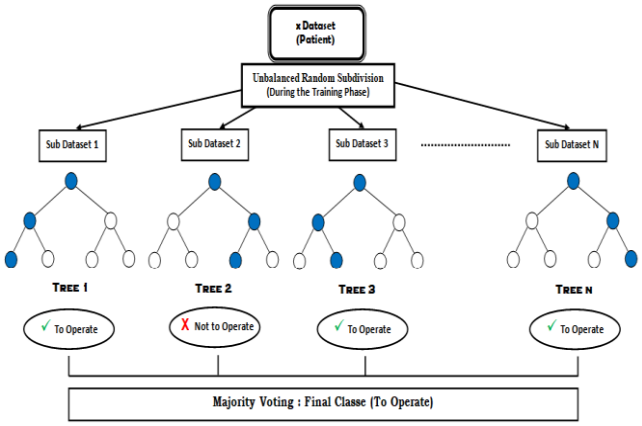

Fig. 6 Detailed Description of the Approach.

V. Case Study

In order to show the feasibility of our approach, we used a database (Dataset) approved by the Institutional Review Board of Seoul National University Hospital (H-1408-101-605). The study has also been registered on clinicaltrials.gov (NCT02914444) [10]. Data collection was carried out in accordance with the relevant guidelines and regulations of the institution's ethics committee. For simplicity, we have generated a small part of this database (see TABLE I), which has been pre-processed to simulate our approach:

TABLE I
Used part of the Dataset

ID	Age	Sex	Preop Hypertension	Body temperature	Preop ECG	Preop Hemoglobin	Preop Glucose	Operate
1	63	M	120/80 mmHg	36.1°C	Worrying	13.5 g/dL	100 mg/dL	Yes
2	34	M	100/65 mmHg	37.1°C	Normal	13.8 g/dL	130 mg/dL	Yes
3	52	F	Elevated	37.8°C	Worrying	10.0 g/dL	70 mg/dL	No
4	77	F	115/70 mmHg	38.1°C	Worrying	14.9 g/dL	85 mg/dL	Yes
5	84	M	Elevated	38.6°C	Normal	17.2 g/dL	115 mg/dL	Yes
6	66	M	118/78 mmHg	36.4°C	Normal	16.3 g/dL	96 mg/dL	Yes
7	81	M	105/65 mmHg	37.5°C	Normal	15.5 g/dL	126 mg/dL	Yes
8	72	F	115/75 mmHg	38.0°C	Worrying	10.5 g/dL	132 mg/dL	No
9	75	F	Stage 1	38.1°C	Normal	12.9 g/dL	119 mg/dL	Yes
10	83	F	Stage 2	37.3°C	Normal	13.2 g/dL	122 mg/dL	Yes
11	51	M	Elevated	36.5°C	Worrying	13.9 g/dL	114 mg/dL	No
12	65	M	Elevated	39.1°C	Normal	14.1 g/dL	128 mg/dL	Yes
13	80	F	117/68 mmHg	36.9°C	Worrying	11.0 g/dL	90 mg/dL	No
14	20	F	Stage 2	37.7°C	Normal	15.4 g/dL	75 mg/dL	Yes
15	86	M	Elevated	38.8°C	Normal	15.5 g/dL	88 mg/dL	Yes
16	79	F	Stage 1	36.8°C	Normal	12.9 g/dL	121 mg/dL	Yes
17	93	M	110/70 mmHg	40.1°C	Worrying	13.5 g/dL	125 mg/dL	No
18	17	F	Stage 2	39.2°C	Normal	15.0 g/dL	73 mg/dL	Yes
19	69	F	119/79 mmHg	37.5°C	Worrying	11.3 g/dL	80 mg/dL	No
20	23	M	Elevated	39.7°C	Normal	13.7 g/dL	95 mg/dL	Yes

We notice that each attribute has multiple values, so we'll use a binary recursive partitioning algorithm, such as CART, to build each decision tree in our random forest. For this, we need to discretize our continuous attributes by creating thresholds (see TABLE II), and this to divide the data into two binary groups. This will allow us to build a binary tree using the discrete values of each attribute.

TABLE II
Attribute Threshold Repository

Age		Hypertention		Body temperature		ECG		Hemoglobin		Glucose	
young	Old	Normal	Worrying	Normal	Worrying	Normal	Worrying	Normal	Worrying	Normal	Worrying
< 50 Y	> 50 Y	Less than 120/80 mmHg	More than 120/80 mmHg (Elevated, Stage 1, Stage 2)	36.0°C - 38.1°C	> 38.1°C	Between 60 and 100bpm	Other	Men : 13.5 -17.5 g/dL Women : 12.9 - 15.5 g/dL.	10.0 - 12.9 g/dL	70 - 100 mg/dL	> 126 mg/dL

We get the following result (see TABLE III):

TABLE III
ATTRIBUTE THRESHOLD REPOSITORY WITH BINARY VALUES

ID	Age	Sex	Preop Hypertension	Body temperature	Preop ECG	Preop Hemoglobin	Preop Glucose	Operate
1	O	M	Normal	Normal	Worrying	Normal	Normal	Yes
2	Y	M	Normal	Normal	Normal	Normal	Worrying	Yes
3	O	F	Worrying	Normal	Worrying	Worrying	Normal	No
4	O	F	Normal	Normal	Worrying	Normal	Normal	Yes
5	O	M	Worrying	Normal	Normal	Normal	Worrying	Yes
6	O	M	Normal	Normal	Normal	Normal	Normal	Yes
7	O	M	Normal	Normal	Normal	Normal	Worrying	Yes
8	O	F	Normal	Normal	Worrying	Worrying	Worrying	No
9	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
10	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
11	O	M	Worrying	Normal	Worrying	Normal	Worrying	No
12	O	M	Worrying	Worrying	Normal	Normal	Worrying	Yes
13	O	F	Normal	Normal	Worrying	Worrying	Normal	No
14	Y	F	Worrying	Normal	Normal	Normal	Normal	Yes
15	O	M	Worrying	Worrying	Normal	Normal	Normal	Yes
16	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
17	O	M	Normal	Worrying	Worrying	Normal	Worrying	No
18	Y	F	Worrying	Worrying	Normal	Normal	Normal	Yes
19	O	F	Normal	Normal	Worrying	Worrying	Normal	No
20	Y	M	Worrying	Worrying	Normal	Normal	Normal	Yes

Next, a step of dividing the data set into training and test sets is carried out. This step allows us to evaluate the performance of our model, through its implementation in Python version 3.9.16 with Google Colab. The training set is used to tune the individual decision trees in the random forest, while the test set is used to assess the accuracy of the predictions made by the model.

TABLE IV
DIVISION OF THE DATASET INTO TRAINING DATA, AND TEST DATA

Training sets

ID	Age	Sex	Preop Hypertension	Body temperature	Preop ECG	Preop Hemoglobin	Preop Glucose	Operate
1	O	M	Normal	Normal	Worrying	Normal	Normal	Yes
2	Y	M	Normal	Normal	Normal	Normal	Worrying	Yes
3	O	F	Worrying	Normal	Worrying	Worrying	Normal	No
4	O	F	Normal	Normal	Worrying	Normal	Normal	Yes
5	O	M	Worrying	Normal	Normal	Normal	Worrying	Yes
6	O	M	Normal	Normal	Normal	Normal	Normal	Yes
7	O	M	Normal	Normal	Normal	Normal	Worrying	Yes
8	O	F	Normal	Normal	Worrying	Worrying	Worrying	No
9	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
10	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
11	O	M	Worrying	Normal	Worrying	Normal	Worrying	No
12	O	M	Worrying	Worrying	Normal	Normal	Worrying	Yes
13	O	F	Normal	Normal	Worrying	Worrying	Normal	No
14	Y	F	Worrying	Normal	Normal	Normal	Normal	Yes
15	O	M	Worrying	Worrying	Normal	Normal	Normal	Yes

Testing sets

ID	Age	Sex	Preop Hypertension	Body temperature	Preop ECG	Preop Hemoglobin	Preop Glucose	Operate
16	O	F	Worrying	Normal	Normal	Normal	Worrying	Yes
17	O	M	Normal	Worrying	Worrying	Normal	Worrying	No
18	Y	F	Worrying	Worrying	Normal	Normal	Normal	Yes
19	O	F	Normal	Normal	Worrying	Worrying	Normal	No
20	Y	M	Worrying	Worrying	Normal	Normal	Normal	Yes

In random forest, the first tree is constructed using a combination of bagging and sampling. Bagging stands for bootstrap aggregation, which involves randomly sampling the training data with replacement to create multiple new datasets. Each of these datasets is the same size as the original dataset, but they contain different observations due to replacement sampling [7]. After creating the new datasets, a decision tree is built for each of them using a random subset of the features. The number of features to use is specified by a hyper parameter. This is called feature sampling. The tree is built by recursively dividing the data into smaller and smaller subsets based on the values of the selected features, until a stopping criterion is satisfied [2].

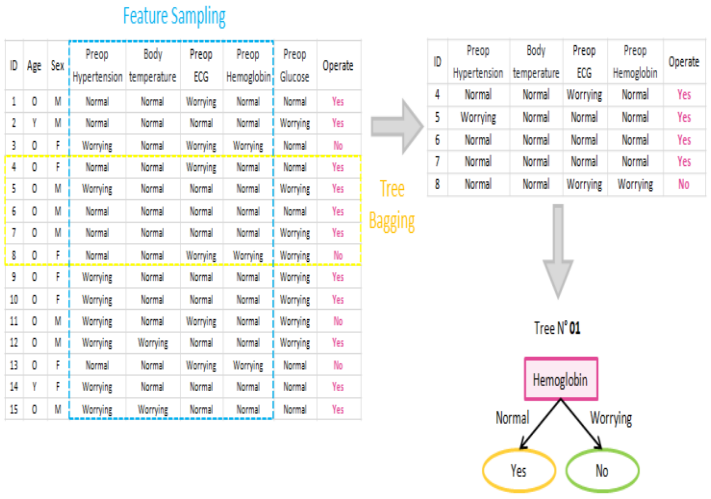

Fig. 7 Construction of the first tree in the Random Forest algorithm

After generating several decision trees, the result is a set of decision trees, each of which is constructed using a different set of data and a random subset of the features. As shown in the figure below:

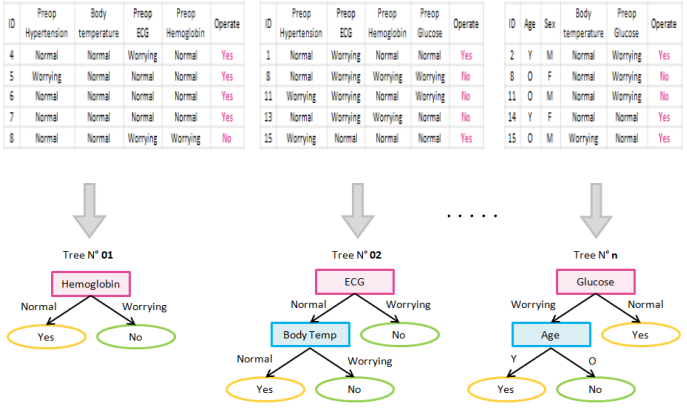

Fig. 8 Generation of several trees by the Random Forest algorithm

To make a prediction about a new instance, the prediction for each individual tree is computed, and the final prediction is obtained through a majority vote (see Fig. 9).

TABLE V
NEW INSTANCE TO TEST

Id	Age	Sex	Preop Hypertension	Body temperature	Preop ECG	Preop Hemoglobin	Preop Glucose	Operate
X	O	M	Normal	Worrying	Worrying	Normal	Worrying	?

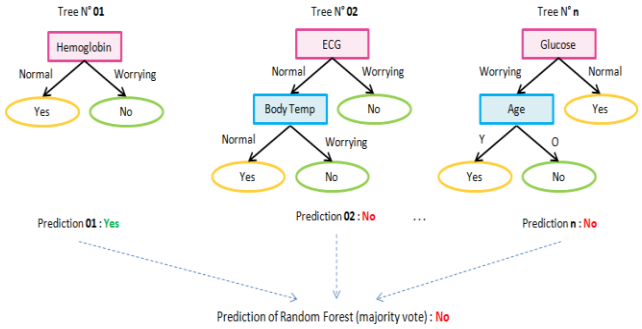

Fig. 9 Prediction of a new instance

VI. CONCLUSION AND PERSPECTIVES

Through this paper we have presented a new approach to a problem that has largely been the preoccupation and major concern of managers in the surgical field; This is the preoperative phase, a deterministic phase in relation to the importance of patient care, as well as in relation to the proper management of the operating room. So the idea guided us to propose an intelligent decision support system for the surgical preoperative phase, based on Machine Learning models, and specifically on the Random Forest algorithm, while exploiting data preoperative predictions drawn from a Dataset. Therefore, we presented our considered intelligent system approach for preoperative surgical decision-making, an approach that was supported through a case study. As prospects, we believe that the integration of this new idea in the surgical field offers an advanced and simple solution, and will make a new revolution, claiming to be able to apply and adopt it in larger Datasets. While trying to prove the effectiveness of our model by comparing it to other models, namely: K-NN, SVM... etc.

REFERENCES

[1] Al Bataineh, A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing*, A. (2019). *Vol 9*, 248-254.

[2] Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow", March 13, 2017.

[3] Babar, A. H., & Mahoto, Comparative Analysis of Classification Models for Healthcare Data Analysis. *International Journal of Computer and Information Technology*, (2018). *Vol 7*(4), 170-175.

[4] BOU SALEH Bilal. Distributed Artificial Intelligence approach for reactive planning and assistance in the conduct of the hospital operating theater process. Doctoral thesis. Presented and defended at "Belfort", on "December 19, 2019".

[5] E. George, E. Flagg, K. Chang, H.X. Bai, H.J. Aerts, M. Vallières, D.A. Reardon and R.Y. Huang Radiomics-Based Machine Learning for Outcome Prediction in a Multicenter Phase II Study of Programmed Death-Ligand 1 Inhibition Immunotherapy for Glioblastoma. American Journal of Neuroradiology May 2022, 43 (5) 675-681.

[6] Ferreira AM, Santos LI, Sabino EC, Ribeiro ALP, Oliveira-da Silva LCd, Damasceno RF, et al. Two-year death prediction models among patients with Chagas Disease using machine learning-based methods. PLoS Negl Trop Dis 16(4): e0010356. (2022).

[7] Hong W, Lu Y, Zhou X, Jin S, Pan J, Lin Q, Yang S, Basharat Z, Zippi M and Goyal H (2022) Usefulness of Random Forest Algorithm in Predicting Severe Acute Pancreatitis. Front. Cell. Infect. Microbiol. 12:893294. doi: 10.3389/fcimb.2022.893294.

[8] Hongcheng Wei , Jie Sun , Wenqi Shan , Wenwen Xiao , Bingqian Wang , Xuan Ma , Weiyue Hu , Xinru Wang , Yankai Xia. Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus. Available online 29 September 2021, Version of Record 3 October 2021.

[9] Hyer JM, White S, Cloyd J, Dillhoff M, Tsung A, Pawlik TM, et al. Can we improve prediction of adverse surgical outcomes ? Development of a surgical complexity score using a novel machine learning technique. J Am Coll Surg. 2020;230(1):43-52.e1.

[10] Lee, HC., Park, Y., Yoon, S.B. et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci Data 9, 279 (2022).

[11] Lee SH, Lee CH, Hwang SH, Kang DH. A Machine Learning-Based Prognostic Model for the Prediction of Early Death After Traumatic Brain Injury: Comparison with the Corticosteroid Randomization After Significant Head Injury (CRASH) Model. World Neurosurg. 2022 Oct;166:e125-e134. doi: 10.1016/j.wneu.2022.06.130. Epub 2022 Jul 3. PMID: 35787963.

[12] Liu, Y.H., Jin, J. & Liu, Y.J. Machine learning–based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Support Care Cancer 30, 2507–2513 (2022).

[13] Melstrom LG, Rodin AS, Rossi LA, Fu P, Fong Y, Sun V. Patient generated health data and electronic health record integration in oncologic surgery: A call for artificial intelligence and machine learning. J Surg Oncol. 2021;123(1):52-60.

[14] Michael B. Wallace Prateek Sharma Pradeep Bhandari James. Impact of Artificial Intelligence on Miss Rate of Colorectal Neoplasia. Published : March 15, 2022. ORIGINAL RESEARCH FULL REPORT: ARTIFICIAL INTELLIGENCE| VOLUME 163, ISSUE 1, P295-304.E5, JULY 2022

[15] Robin Genuer and Jean-Michel Poggi "CART Trees and Random Forests, Importance and Selection of Variables", January 2017.

[16] Shanthababu Pandian "Understanding machine learning and its end-to-end process", 2020.

[17] Tighe D, Lewis-Morris T, Freitas A. Machine learning methods applied to audit of surgical outcomes after treatment for cancer of the head and neck. Br J Oral Maxillofac Surg. 2019;57(8):771-7.

[18] Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, Han B, Wang H, Duan X and Sun T (2021) A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. Front. Public Health 9:754348. doi: 10.3389/fpubh.2021.754348

[19] Yalong Zhang, Zunni Zhang, Liuxiang Wei and Shujing Wei. Construction and validation of nomograms combined with novel machine learning algorithms to predict early death of patients with metastatic colorectal cancer. Front. Public Health, 20 December 2022 Sec. Aging and Public Health Volume 10 – 2022.

[20] Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, Yu W, Yan J. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020 Mar 23;10(1):5245. doi: 10.1038/s41598-020-62133-5. PMID: 32251324; PMCID: PMC7090086.

Leveraging Machine Learning for Accurate Malware Traffic Detection

Elshan Baghirov

Ministry of Science and Education of the Republic of Azerbaijan
Institute of Information Technology
Azerbaijan, Baku
elsenbagirov1995@gmail.com

***Abstract*— This paper proposes a novel approach to accurately detect malware traffic using machine learning techniques. The increasing prevalence of malware and its sophisticated nature has made it difficult to identify and prevent attacks. In this work, we applied various machine learning algorithms to detect malware traffic by analyzing network traffic patterns. Our proposed method outperforms existing solutions and achieves high accuracy in detecting both known and unknown malware samples. The results demonstrate the effectiveness of machine learning in improving the accuracy of malware detection, and highlight the potential of this approach in enhancing cybersecurity measures.**

***Keywords*— *malware traffic, machine learning, malware detection, cybersecurity, network* traffic**

I. Introduction

Malware is a pervasive threat to computer security that can cause significant harm to individuals, organizations, and governments. Cybercriminals use malware to steal sensitive information, launch DDoS attacks, and compromise systems. One of the critical challenges in detecting malware is the ability to distinguish between legitimate and malicious network traffic. Traditional signature-based methods can be easily evaded by attackers, making them less effective in identifying sophisticated malware. TLS poses a challenge because it encrypts the traffic, making it difficult to inspect and anlyze. However, there are several approaches to detecting malware traffic that uses TLS [1]. Machine learning algorithms have emerged as a promising solution to this problem, as they can learn from large amounts of data and identify patterns that are indicative of malicious traffic.

While machine learning has shown significant promise in malware detection, there are several challenges to its implementation. One challenge is the identification of suitable features to detect malware traffic, as it is essential to provide relevant information to the machine learning algorithm. Another challenge is the need to train machine learning models on high-quality data that accurately represent the malware traffic. Finally, machine learning models must be designed to be computationally efficient to run in real-time and scalable to handle large amounts of traffic.

Therefore, in this research, we aim to leverage machine learning techniques for accurate malware traffic detection. Specifically, we will explore various ML algorithms and evaluate their effectiveness in detecting malware traffic in real-world network environments. We will also investigate the use of feature selection and dimensionality reduction techniques to improve the accuracy and efficiency of our models. We will conduct experiments using public datasets such as the Malware Traffic Analysis Dataset (MTA-KDD'19) [1] which contains a broad range of malware traffic samples.

The results of this research have the potential to significantly improve malware detection capabilities and help organizations better protect their networks and data. By leveraging the power of machine learning, we can stay ahead of the evolving threat landscape and ensure a more secure digital future.

II. Related Works

Several previous works have explored the use of machine learning techniques for malware traffic detection. In this section, we review some of the most relevant studies.

Minghui G. et al. [2] presented a two-tiered system for detecting anomalies in network traffic using deep neural networks and association analysis. The effectiveness of different neural network models was evaluated using publicly available datasets, and DNN-4 was chosen as the most effective for identifying malicious traffic. However, the deep neural network's architecture can still be improved, and maintaining a high precision rate while increasing the recall rate is a challenging task that could be explored.

Wing Z. et al. [3] conducted an analysis of data, processed it, and merged it from five different sources with the intention of creating a comprehensive and equitable dataset that could be used for further research in their field. In addition, they also executed and compared 10 encrypted malicious traffic detection algorithms based on this dataset. The primary concern is the absence of a thorough, well-balanced, authentic, and persuasive public dataset in the field of identifying encrypted malicious traffic.

Marin G. et al. [4] presented DeepMAL, a deep learning model that can learn the fundamental patterns of malicious traffic without relying on manually created features by experts to make approach flexible and generic. Using real network measurements, publicly available though the Stratosphere IPS Project of the CTU University of Prague in Czech Republic authors demonstrated that utilizing Raw Flows as the input for deep learning models produces significantly superior outcomes in comparison to using Raw Packets. However, the complexity of the deep learning models used in the study may make it difficult to interpret the results and identify any errors or inaccuracies. Also, the use of deep learning models requires

significant computational resources, which may make it challenging to implement in certain environments.

III. Methodology

To detect and analyze malware traffic, it is utilized a combination of behavior-based detection, and machine learning-based detection techniques. The following steps were taken to conduct the research:

1. *Dataset preparation*: The collected data was preprocessed to remove any noise or irrelevant data, handle missing data, scaling etc.
2. *Machine learning-based detection*: We used mostly used machine learning algorithms to analyze the preprocessed data and identify patterns that are associated with malware traffic.
3. *Validation and testing*: Accuracy and effectiveness of our detection methods was validated by comparing the results with ground-truth data and by conducting testing using a separate dataset.
4. *Optimization*: Detection methods was optimized by tuning the parameters of the algorithms and by incorporating feedback from the validation and testing stages.

IV. Experiments

A. Dataset. The dataset used in this study is a publicly available dataset called the MTA-KDD'19 dataset [5]. It is a refined and updated dataset designed for training and evaluating machine learning-based malware traffic analysis algorithms. It was created by starting with the largest databases of network traffic captures available online and preprocessing it to remove noise and handle missing data. The resulting dataset is unbiased and specifically tailored to machine learning algorithms, and the entire process can be run automatically to keep it updated. Dataset has approximately 30 000 benign and 34 000 malware samples with 33 features.

B. Training information. As a learning model, we have used several classification algorithms: Random Forest, Logistic Regression, LightGBM, Decision Tree, KNN, SVM, SGD. Methods are evaluated using different metrics. Results of classification are shown in table 1. Random Forest achieved a high score than others. Results demonstrate the effectiveness of our machine learning approach to malware traffic detection.

Table 1. Experiments results

Methods	F1-score	Precison	Recall
Random Forest	***0.998***	***0.998***	***0.998***
LightGBM	0.998	0.999	0.996
Decision Tree	0.964	0.978	0.949
Logistic regression	0.844	0.836	0.852
Gradient Boosting	0.993	0.991	0.995
K-nearest neighbors	0.979	0.991	0.968
SGD	0.838	0.844	0.832
SVM	0.984	0.986	0.981

Results

In conclusion, the findings of this research paper suggest that malware traffic detection is a complex and evolving field that requires a multifaceted approach. Network monitoring, machine learning, and intrusion detection systems all have a role to play in identifying and preventing the spread of malware.

To improve malware traffic detection, further research is needed to develop more sophisticated algorithms and techniques that can detect new and unknown malware effectively. Additionally, organizations need to invest in regular updates to software and security protocols to keep their networks safe from malware infections. Ultimately, effective malware traffic detection is critical in preventing significant damage to networks and their users.

References

[1] Jiyuan L. et al. MalDetect: A Structure of Encrypted Malware Traffic Detection, Computers, Materials & Continua, vol.60, no.2, pp.721-739, 2019.

[2] Minghui G. et al. Malicious Network Traffic Detection Based on Deep Neural Networks and Association Analysis, Journal of Sensors, 2020, pp.1-14.

[3] Wang Z. et al. Machine Learning for Encrypted Malicious Traffic Detection: Approaches, Datasets and Comparative Study, arXiv:2203.09332v1, 2022, pp.1-21.

[4] Marin G. et al. DeepMAL - Deep Learning Models for Malware Traffic Detection and Classification, Proceedings of the 3rd International Data Science Conference – iDSC2020, 2021, pp.105-112.

[5] Letteri I. et al., MTA-KDD'19: Dataset for Malware Traffic Detection, Proceedings of the Fourth Italian Conference on Cyber Security, Ancona, Italy, February 4th to 7th, 2020, pp.153-165.