BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 05/07 pp33–46
Volume 4, Number 1, March 2021
DOI: 10.26599/BDMA.2020.9020023

# Improvement in Automated Diagnosis of Soft Tissues Tumors Using Machine Learning

El Arbi Abdellaoui Alaoui, Stéphane Cédric Koumetio Tekouabou*, Sri Hartini, Zuherman Rustam, Hassan Silkan, and Said Agoujil

**Abstract:** Soft Tissue Tumors (STT) are a form of sarcoma found in tissues that connect, support, and surround body structures. Because of their shallow frequency in the body and their great diversity, they appear to be heterogeneous when observed through Magnetic Resonance Imaging (MRI). They are easily confused with other diseases such as fibroadenoma mammae, lymphadenopathy, and struma nodosa, and these diagnostic errors have a considerable detrimental effect on the medical treatment process of patients. Researchers have proposed several machine learning models to classify tumors, but none have adequately addressed this misdiagnosis problem. Also, similar studies that have proposed models for evaluation of such tumors mostly do not consider the heterogeneity and the size of the data. Therefore, we propose a machine learning-based approach which combines a new technique of preprocessing the data for features transformation, resampling techniques to eliminate the bias and the deviation of instability and performing classifier tests based on the Support Vector Machine (SVM) and Decision Tree (DT) algorithms. The tests carried out on dataset collected in Nur Hidayah Hospital of Yogyakarta in Indonesia show a great improvement compared to previous studies. These results confirm that machine learning methods could provide efficient and effective tools to reinforce the automatic decision-making processes of STT diagnostics.

**Key words:** classification; soft tissues tumours; preprocessing techniques; Support Vector Machine (SVM); Decision Tree (DT); machine learning; predictive diagnosis

## 1 Introduction

The term “soft tissue” refers to tissues that support, connect, or surround other structures and organs in the body such as fat, muscles, blood vessels, deep cutaneous tissues, nerves, and tissues surrounding the joints (synovial tissue)[1]. As the name suggests, these are sensitive tissues that can be affected by several infections, including tumors that can develop almost anywhere in the human body. The malignant types of these tumors, also known as Soft Tissue Sarcomas (STS), are grouped together because they share many microscopic features, exhibit the same symptoms, and are almost similarly treated[1,2]. Yet, effective diagnosis of Soft Tissues Tumors (STT) is still a big challenge owing to the difficulty in detecting these cancers. Several techniques have therefore been developed to strengthen the detection of such cancers, including Magnetic Resonance Imaging (MRI) analysis. MRI is currently considered the standard diagnostic tool for the detection and classification of STT[3] with well characterized

• El Arbi Abdellaoui Alaoui and Said Agoujil are with the Department of Computer Sciences, Faculty of Sciences and Technologies, My Ismail University, Errachidia 52000, Morocco. E-mail: abdellaoui.e@gmail.com; agoujil@gmail.com.
• Stéphane Cédric Koumetio Tekouabou and Hassan Silkan are with the Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, El Jadida 24000, Morocco. E-mail: ctekouaboukoumetio@gmail.com; silkan.h@ucd.ac.ma.
• Sri Hartini and Zuherman Rustam are with the Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia, E-mail: {sri.hartini, rustam}@sci.ui.ac.id.
∗ To whom correspondence should be addressed.
Manuscript received: 2020-08-01; accepted: 2020-09-23

biological properties such as cellular origins and tumour specimens[4] used to distinguish tumors. MRI can be used to analyze textural characteristics or other less characterized tumor characteristics (average MRI signal intensity, shape of tumor boundaries) for several reasons: (1) ease of computation textural characteristics, (2) wide correlation of textural characteristics to tumor pathology[5,6], and (3) robustness to changes in MRI acquisition parameters such as changes in the resolution of the tumor image and the corruption of the MRI image due to heterogeneity of the magnetic field[2]. Such magnetic field heterogeneity in MRI makes it difficult to perceive the texture in certain malignant tumors and humans have a limited capacity to perceive and discriminate these textures as well[7]. Hence there is an increasing use of Machine Learning (ML) algorithms to analyze MRI images more effectively and automatically detect cancers. It has become an essential tool for modern medicine today and has been strengthened by predictive automatic learning algorithms that improve the diagnostic performance of existing expert systems[3]. Among these many applications, we have developed a machine learning-based technique for the automotive detection and diagnosis of tumors such as STT.

STT are malignant tumors that develop within tissues such as fat, muscles, nerves, fibrous tissues, and blood vessels. Because of their low frequency and the difficulty physicians have interpreting results, these challenges have prevented the development of new therapeutic agents. In addition, the inconsistent MRI images make it difficult for physicians to determine an effective treatment[8]. Besides, STT can easily be confused with other diseases such as fibroadenoma mammae, lymphadenopathy, and struma nodosa. This diagnostic failure has a significant impact on the patient treatment process. According to the theory mentioned by Karanian and Coindre[9], there are four categories of connective tumor evolution benign lesions, tumors with local potential, tumors with low metastatic potential, and sarcomas. When a molecular anomaly of an entity has been identified, the definition of this entity, which is both histological and molecular, is obtained[9]. The current challenge is, therefore, how to effectively use the characteristics of these anomalies for better targeted therapy for STT.

The predictive detection of STT is reinforced by the use of classification techniques and necessary to avoid delays in diagnosing the patient and optimizing their treatment. That is why Nur Hidayah Hospital in Yogyakarta, Indonesia has been interested in predicting whether a patient is correctly diagnosed with the STT or non-STT in order to provide effective treatment. To do this, we analyze a dataset consisting of 50 patients who were diagnosed with the STT and 25 patients wrongly diagnosed with the STT (non-STT). Additional criteria included in the dataset are (1) all patients had completed the Complete Blood Count (CBC) and blood clotting tests; (2) the result of their Total Protein & Alburnin/Globumin (AGS-AS) antigen test is negative. Other patients that were not diagnosed with the STT, although could be mistaken as the STT, had other diseases such as fibroadenoma mammae, lymphadenopathy, and struma nodosa[10]. The main objective of our study is to build and test equipped machine learning-based models that enable users to thoroughly analyze patient data and predictively separate STT from non-STT. The method we built strives to correct the shortcomings encountered in previous work associated with the predictive detection of STT[10, 11].

The rest of the paper is organized as follows: In Section 2, we present the qualities of STT and the challenges of applying ML techniques for STT classification. In Section 3, we provide a systematic and detailed process for constructing and evaluating an automatic learning classifier based on the Support Vector Machine (SVM) and Decision Tree (DT) algorithms that can be used for practical applications with almost the same performance. Our research in this section will also answer some questions related to the problem of predictive analysis of STT, such as the complexity of classifiers, the effect of the learning dataset size on the behavior of the classifier, and the appropriate size of the training data that can be used to train a model of classifier and obtain excellent generalization performance on invisible data. Section 4 evaluates and analyzes the obtained results that we discuss in Section 5, and Section 6 concludes our work.

## 2 STT Classification

Soft tissue is the tissue that surrounds, supports, and connects organs and other parts of the body and gives shape and structure to the body, protects organs, circulates liquids like blood from one body part to another, and stores energy. It is found throughout the body. There are many types of soft tissues, including fat, muscle, fibrous tissue, blood vessels, lymphatic vessels, and nerves. These soft tissues can be affected by many kinds of diseases, including tumors. STT are malignant tumors that develop within soft tissues like fat, muscle, nerves, fibrous tissue, and blood vessels[8].

STT comprise roughly less than 1% of adult cancers with a lifetime risk of development estimated at 0.33%[1, 12]. Therefore, STT have preserved for decades an aura of mystery linked to their diagnosis as deemed unmanageable. They were first characterized by their traditionally unfortunate prognosis, and limited glimmers of hope for recovery[1]. Indeed, the diagnosis of STT has been marked by profound advancements in investigation methods with the intervention of cytogenetics and molecular biology. Furthermore, the discovery of recurrent anomalies in other areas of pathology has led to reconsideration of traditional histological frameworks. These improvements gave birth in 2002 to a new version of the World Health Organization (WHO) classification of soft tissue tumors[13, 14], taking into account genetic and molecular data. This edition organized STT types according to the following categories: adipocytic, fibroblastic/myofibroblastic, fibrohistiocytic, smooth muscle, pericytic/perivascular, skeletal muscle, vascular, chondro-osseous tumors, and the group called “uncertain differentiation tumors”. Eleven years after this third edition, the fourth edition of the classification of STT by the WHO was published in February 2013 and was made according to the type of tumor, and the morphological, immunohistochemical, and genetic characteristics[9, 15]. The 2013 edition also includes chapters on gastrointestinal stromal tumors and nerve sheath tumors and a newly introduced section for undifferentiated/unclassified sarcomas[9]. This WHO classification has made it possible to diagnose specific types of cancer more effectively in order to give patients the best treatments. Currently, this diagnosis is making increasing use of artificial intelligence, which uses ML algorithms for better diagnosis of tumors. The following section will discuss the challenges of using ML for the classification of STT.

## 2.1 Immunohistochemistry for STT diagnosis

Immunohistochemistry remains an easily accessible and inexpensive technique and an essential step in the diagnostic process. It is based on the principle of the antigen-antibody reaction[13]. Immunohistochemistry makes it possible to highlight several types of antigens potentially present at certain levels within cells such as markers of cell differentiation and anti-oncogenes (protein INI-1 in rhabdoid tumors), and markers of cell proliferation (Ki-67). Correlated to the morphology of tumor, the immunohistochemical markers of differentiation make it possible in many cases to determine the line of differentiation of the cells and determine a diagnosis. In some types of STT, immunohistochemistry can be used as an indirect witness to a chromosomal abnormality[16]. Immunostaining with anti-WT1 antibodies (clone C19) (Figs. 1c and 1d), anti-Fli-1 (Figs. 1a and 1b), or anti-mdm2 is useful in the diagnosis of desmoplastic small cell tumor well-differentiated and undifferentiated rounds, Ewing's sarcoma, and liposarcomas, respectively.

## 2.2 STT versus non-STT

STT, especially high-grade STT, do not always have specific morphological characteristics that allow them to be recognized and differentiated from non-STT. Before making a diagnosis of STT, the pathologist must imperatively rule out other malignant tumors such as

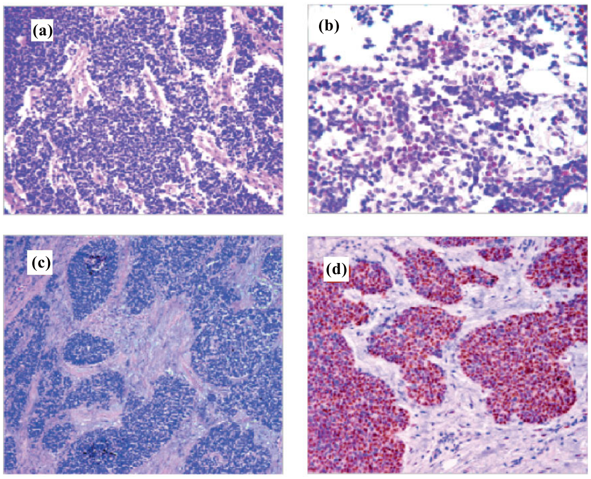


**Fig. 1 (a) Peritoneal nodule composed of tumor masses of variable size, enclosed in a large desmoplastic stroma. The tumor population is characterized by small round monomorphic cells with a high nucleocytoplasmic ratio. Some masses are centered by necrosis[16]. (b) Nuclear labelling of tumour cells with the antibody directed against the COOH- (carboxylic acid) end of the wilm's tumor 1 (WT-1) protein. It is an indirect reflection of the t(11; 22) (q24; q12) translocation involving the Ewing Sarcoma (EWS) genes on chromosome 22 and the WT-1 gene on chromosome 11[16–18]. (c) Tumor proliferation of diffuse architecture composed of small round monomorphic cells with a high nucleocytoplasmic ratio[16]. (d) Nuclear labeling of tumor cells by the antibody directed against the friend leukemia integration 1 (FLI-1) transcription factor. It is an indirect reflection of a t(11; 22) (q24; q12) translocation involving the EWS gene on chromosome 22 and the Fli-1 gene on chromosome 11[16–18].**

a sarcomatoid carcinoma, melanoma, or lymphoma. Those tumors are much more prevalent than STT, and require different therapeutic managements than STT. Immunohistochemistry is an essential step in the diagnostic process, and in some cases, molecular analysis in search of specific genetic anomalies of certain sarcomas is necessary[16].

### 2.3 Challenge of ML for soft tissues tumor classification

The automatic classification of STT is a challenging problem due to their very varied morphological appearances. The size and the tumors can vary greatly from one patient to another[1]. In addition to this, the tumor boundaries are generally unclear, non-uniform, and irregular with discontinuities, which poses a significant challenge, especially when traditional diagnosis methods depend on clear tumor boundaries. Additionally, complex MRI data for tumors from clinical analyses or synthetic databases are difficult for physicians to interpret[2]. The MRI devices and protocols used for imaging pricing can vary from one scan to another, imposing intensity bias and other variations on each different image slice in the dataset. The fact that the STT are very similar to one another, makes their classification very difficult. Indeed, after the WHO classification in 2013, 20% of these tumors remained undifferentiated and classed due to their different morphological appearances[9, 15]. The need for several methods to effectively segment the tumor sub-regions adds to their complexity. All these factors, along with computer advances, have made computer-aided diagnostic methods more efficient than expert systems centered around a doctor[19]. For several decades, these computer-aided diagnostic methods have been increasingly reinforced by the integration of automatic targeting algorithms in order to detect symptomatic behaviors linked to each type of disease to be diagnosed. In this study, we have built a prediction model that can classify STT versus non-STT automatically using the finesse of these algorithms to strengthen the diagnotic systems of soft tissue tumors. The next section will outline the data we used and the detailed process of building our prediction model.

## 3 Material and Method

### 3.1 STT dataset

The STT dataset consists of data from biological analysis of 75 patients, among which 50 patients who were diagnosed with the STT and 25 patients with diseases wrongly diagnosed with the STT. The STT dataset has been obtained from Nur Hidayah Hospital, Yogyakarta, Indonesia. The patients with the STT analyzed in this study were comprised of 24 males and 26 females from 1 to 77 years old and a mean age of 38 years old[10]. The additional criteria used in selecting patients from the dataset include:

(1) All patients had completed a CBC and blood clotting test;

(2) The result of the AGS-AS antigen test is negative;

(3) Other patients that were not diagnosed with STT but with diseases that can be incorrectly diagnosed as STT such as fibroadenoma mammae, lymphadenopathy, and struma nodosa.

The STT dataset is comprised of 20 attributes, including labels, with 15 numerical and five categorical attributes as summarized in Table 1[10]. The dataset contains null attributes that require special preprocessing before classification with ML algorithms, as will be explained in the next section.

### 3.2 Data preprocessing

Our proposed approach for data preprocessing consists of two main steps (see Fig. 2): (1) preprocessing and (2) classification. The preprocessing step transforms different types of features into a specific format reducing the processing time and improving the classification performance while remaining very stable against variables scales by dealing with each attribute separately. To clearly illustrate these different steps, let us consider an example of instances from Table 1 which contains some common features available in major customers datasets.

#### 3.2.1 Features transformation

At this preprocessing step, the raw data from the various tests and information sources are collected and transformed into a numerical format so that the SVM or DT classifiers can be applied. For this, all categorical attributes (blood type, AGS-AS, gender, and disease) were first classified in nominal type (blood type), binomial (AGS-AS, gender, and disease), and missing values, then converted to numerical format in the following way depending on the type[20, 21]:

► For Boolean features such as $f_3$: Given this type of feature, we have only two possibilities: yes or no (1/0), male or female (1/0), or positive/negative(1/0).

► For nominal features (example of "blood type"): The particularity of this approach is how it deals with

**Table 1 Detail about soft tissues tumor dataset.**

| No. | Attribute | Description | Type |
|---|---|---|---|
| 1 | ID | ID of the patient | Numerical |
| 2 | Age | Age of the patient in years | Numerical |
| 3 | Gender | Man or woman | Numerical |
| 4 | Kind of disease | If the patient is diagnosed with STT or not | Categorical |
| 5 | WBC | Number of white blood cells in thousands per microliter of blood | Numerical |
| 6 | RBC | Number of red blood cells in millions per microliter of blood | Numerical |
| 7 | HGB | Number of hemoglobin in grams per deciliter of blood | Numerical |
| 8 | HCT | Hematocrit (the volume percentage of red blood cells in blood) | Numerical |
| 9 | MCV | Average volume of a red blood cell in femtoliter of blood | Numerical |
| 10 | MCH | Average mass of hemoglobin per red blood cell in picograms per littro | Numerical |
| 11 | MCHC | Concentration of hemoglobin in a red blood cell in grams per deciliter of blood | Numerical |
| 12 | PLT | Number of platelet count in thousands per microliter of blood | Numerical |
| 13 | Lymphocytes (%) | Percentage of lymphocytes in blood | Numerical |
| 14 | Monocytes (%) | Percentage of monocytes in blood | Numerical |
| 15 | Neutrophils (%) | Percentage of neutrophils in blood | Numerical |
| 16 | Blood type | Blood type of the patient | Categorical |
| 17 | Clotting-time | Amount of time required for a sample of blood to clot in minutes | Numerical |
| 18 | Bleeding time | Amount of time needed for bleeding to stop in minutes | Numerical |
| 19 | AGS-AG | Total protein & albumin/globumin | Categorical |
| 20 | Blood glucose | Level of glucose in the blood | Numerical |

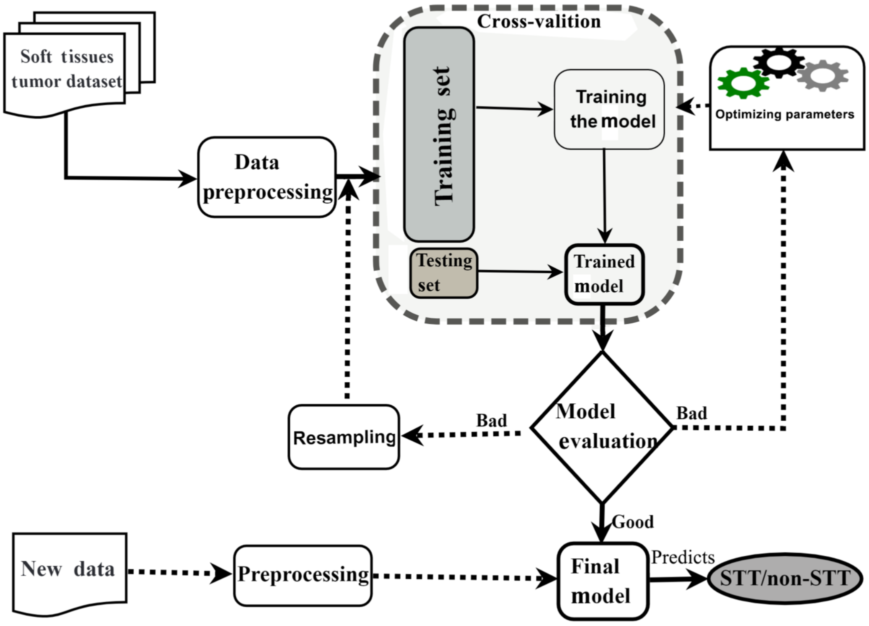


**Fig. 2 Flowchart of machine learning-based system for the automatic discrimination of soft tissue sarcomas.**

nominal features ("blood type" in Table 1). Each value of the nominal features $V_j$ for the example $i$ ($V_{ij}$) can belong to class $q$ ($\mathrm{CL}_q$) if its maximum class frequency is reached for this class in the training set; this frequency is given by Formula (1):

$$V(\mathrm{CL}_q)_{ij} = \begin{cases} 1, & \text{if Max } \dfrac{n_q(V_{ij})}{N_q}; \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $n_q(V_{ij})$ is the number of $V_{ij}$ variables $j$ in the class $\mathrm{CL}_q$ and $N_q$ is the total number of class $q$. Value 1 means that this attribute is more favorable to the class $q$ and will be placed in $\mathrm{CL}_q$ for this variable in the reduced table, and so on for all attributes of nominal variables. So, the nominal attributes take value 1 in this class $q$ and 0 in other classes.

► For missing values: One of the advantages of our approach is that it proposes a function allowing a fast and optimal imputation of the missing values according to the type of features to which they belong[22]. Such function relies on replacing all the missing values $V_{ij}$ of the feature $V_j$ by the average if they are scaled or numerical variable or the mode if they are Boolean or nominal variables inside the class $q$:

$$(V_{ij})_{\text{qmiss}} = \begin{cases} \text{Mode}(V_j)_q, \text{ if Boolean or nominal;} \\ \text{Mean}(V_j)_q, \text{ if numerical} \end{cases} \tag{2}$$

### 3.2.2 Resampling techniques

In general, resampling is necessary when the dataset has been recorded in a given time interval or number of samples and we want to modify this interval or number of samples[2]. The impact of this strategy is also visible on the data distribution, making it more suitable for the classification algorithms. In this case, we want to resize and reorder the data in order to optimize the model's performance. This is done by eliminating the bias and the variance of the algorithms to build a stable and realistic model without overfitting. The process of data resampling is described by Algorithm 1.

## 3.3 Classification

After the preprocessing steps, the next phase of our problem is to apply suitable classification algorithms for discriminating the non-STT from STT. These preprocessing steps have allowed us to correctly clean the data while removing outliers and multicolinearity that could influence the performance of the classifiers. To do this, our approach is based on the optimization of the SVM and DT algorithms that we present briefly in the following sections. Thus, given the data size and type of features, stochastic algorithms would be very suitable for this task. And to do this, the improved DT or SVM algorithms are both simple and powerful to discriminate this type of data. But before applying these algorithms, a statistical analysis should be done in the experiments.

**Algorithm 1 Random data oversampling algorithm**

```
Input: A set of instances $D$;
Output: A set of instances $D'$ such as $D \subseteq D'$;
procedure RANDOM OVERSAMPLING ($D$)
    $D' \leftarrow D$
    $\{\varepsilon_j\}$ is a random sub-set of minority $D$
    for each instance $x_i \in \{\varepsilon_j\}$ in dataset $D$ do
        $D' \leftarrow D' + x_i$ : features $(x_i) \neq$ features $(x_j)$, $\forall (x_j) \in D'$;
    end for
    while no satisfy condition do
        Append $x_i \in \{\varepsilon_j\}$, based on the difference between features $x_j$ and features $\{\varepsilon_j\}$;
    end while
    return $D'$
end procedure
```

### 3.3.1 Approach with SVM classifier

The SVM algorithm is a supervised ML method used for data classification. The primary purpose of using the SVM algorithm is to accurately classify the invisible data by minimizing the misclassification using a decision function[23]. It is done by training the SVM algorithm on the training data and then using the trained model to predict the output class of new data. In our study, an SVM-based approach was applied to classify STT and non-STT tumors. For doing this, the SVM algorithm uses several parameters, such as the kernel. For our study, we chose the nonlinear Radical Basis Function (RBF) kernel which is the most popular kernel used in ML problems[24]. It separates the solution sets which are not linearly separable and is generally presented as follows:

$$K(v_i, v_j) = \exp\left(-\frac{||v_i - v_j||^2}{2\sigma^2}\right) \tag{3}$$

where $v_i$ is the support vector and $v_j$ is the current value of the data. In many representations of the RBF kernel, the factor $\frac{1}{2\sigma^2}$ is replaced by $\gamma$ giving the following equation:

$$K(v_i, v_j) = \exp\left(-\gamma\left(||v_i - v_j||^2\right)\right), \ \gamma > 0 \tag{4}$$

In practice, the main objective of the RBF kernel is to draw the decision boundary that most effectively undermines a set of training data given in two distinct parts. It implies that this approach will be less suitable if the class number is greater than 2. The mathematical representation of this decision boundary function using the RBF kernel is as follows[23]:

$$f(v) = w \cdot K(v_i, v_j) + \lambda, \ w \in v_i, \ \lambda \in \mathbb{R}^n \tag{5}$$

where $w$ is the normal vector to the decision boundary and $\lambda$ is the regularization parameter in $n$-dimensional feature space which is represented by $\mathbb{R}^n$.

To ensure the robustness of $f(x)$, one can for example minimize the norm $||w||$ by adding variable slack $(\xi_i)$ to overcome the occurrence of inequalities in the classification problem[25]. It transforms the decision function represented above and becomes an optimization problem as follows:

$$\min\left(\underbrace{\frac{1}{2}||w||}_{\text{Max. Margin}} - \underbrace{M\sum_{i=1}^{n}\xi_i}_{\text{Min. error}}\right) \tag{6}$$

where $M$ is the additional regularization parameter most known and used to tradeoff between margin to maximize and training error to minimize. To solve Formula (6), we can use the Lagrange multipliers $(\alpha_i, \beta_i)$ which arises to Eq. (7):

$$f(v) = \sum_{i=1}^{n}(\alpha_i - \beta_i)K(v_i, v_j) + \lambda,\ \alpha_i,\ \beta_i \in\ [0,\ M] \tag{7}$$

Regarding the use of the other SVM kernels, in the previous research, Keerthi and Lin[26] demonstrated that when $\sigma^2 \to \infty$, SVM with the RBF kernel and a parameter of penalty $M$ tends to a linear kernel SVM approach with penalty parameter $\frac{M}{2\sigma^2}$. This proof shows that for a better choice of selection parameters, SVM performance with an RBF kernel is as good as using the linear kernel. Lippert and Rifkin[27] also did the same demonstration for a polynomial kernel with a penalty parameter equal to $\left(\frac{1}{2\sigma^2}\right)^{-2d} M$.

### 3.3.2 Approach of DT classifier

DT is a technique for structuring a set of training data in the form of trees consisting of nodes and leaves. Each node represents the test on the given attribute, while the leaf represents the class[21,28,29]. The basic algorithm of induction of the DT is based on the descending recursive method by constructing a DT. The algorithm uses information gain (see Eq. (8)) based on the measurement of entropy as heuristic information and selects a sample of classification attributes that can be called. The attribute becomes the test or decision attribute of the node[11,23].

Consider an attribute $x_j$ of a set of data $T$ with subsets $T_1, T_2, \ldots, T_n$ consisting of cases with distinct known values of attributes $x_j$, then

$$\text{GAIN}(x_j) = \text{Info}(T) - \sum_{i=1}^{n} \frac{|T_i|}{T} \times \text{Info}(T_i) \tag{8}$$

where

$$\text{Info}(T) = -\sum_{k=1}^{\text{NClass}} \frac{\text{freq}(C_k, T)}{T} \times \log_2\left(\frac{\text{freq}(C_k, T)}{T}\right) \tag{9}$$

Info($T$) is the entropy function with $C = \{C_1, C_2, \ldots, C_N\}$ being the classes associated to the data $T$. For the case of C4.5 decision trees, it uses the information gain ratio of the splitting $T_1, T_2, \ldots, T_n$ which is the ratio of information gain to its split information[28,30].

$$\text{Split}(x_j) = -\sum_{i=1}^{n} \frac{|T_i|}{T} \times \log_2\left(\frac{|T_i|}{T}\right) \tag{10}$$

Even though it is fast and straightforward, the algorithm becomes very inefficient for large databases but remains widely used in medical targeting.

## 4 Result and Discussion

In this subsection, we will outline the detailed experimental analysis and measurements of performance results of STT prediction based on SVM and DT algorithms and compare these results to other ML algorithms and the previous research. We will then end by an overall discussion of these results.

### 4.1 Experimental protocol

The first step of the protocol consists in transforming the data and missing values. The whole dataset is then divided into five-folds of which four-fifths are used for training and one-fifth is used for the test. To illustrate the impact of combining of data preprocessing and ML classifiers, we experimented two mean tests. The first test has been done without resampled data and the second test was preceded by training data being resample.

All the experiments were carried out on a Windows platform. All codes were written in the programing language-python 3.7 with the associated free ML library-Scikit-learn[31] library. The computer used is an "Asus"-brand with the following configuration: 8 GB of RAM, Intel core i7 processor, and an NVIDIA Geforce 930M for graphic card.

Our approach is based mainly on two algorithms, SVM and DT. For the SVM classifier, the most important parameter is the kernel for which we chose RBF, which proved to be faster, giving better performance and a regularization parameter $M = 1$ and $\gamma =$ "scale"; while the other parameters were assigned the default. Regarding the DT algorithm experimentation, the parameters were mostly the default parameters of version C4.5, which can be found in the Scikit-learn library of python. In addition to these parameters intrinsic to each function, other parameters such as the split value of train/test data and the optimal size of the data sampled are also necessary as parameters. They will be chosen after cross-validation in the following part in order to construct an unbiased and stable model.

### 4.2 Performance measures

The predictive accuracy rate (Eq. (11)) is the most commonly used metric for optimal performance. However, it is not an effective tool for evaluating models on unbalanced datasets, because it does not indicate how the model correctly classified the minority class instances that are often the targets. Concerning our databases which are unbalanced[32], we will evaluate our approach in terms of other performance measures.

Area Under the Curve (AUC), which is a performance metric generated from the Receiver Operating Characteristic (ROC) curve. The ROC curve is created by plotting the True Positive Rate (TPR) on $y$-axis against the True Negative Rate (TNR) on $x$-axis. It shows the portion misclassified instances and is an ideal performance measurement for imbalance class datasets[33].

$$\text{Accuracy} = \frac{a+b}{b+c+d} \times 100\% \tag{11}$$

$$\text{f1-measure} = \frac{2a}{2a+c+d} \times 100\% \tag{12}$$

where $a$ refers to the set of clients that are correctly predicted $yes$, $b$ refers to the set of clients that are correctly predicted $no$, $c$ is the number of false positives, and $d$ is the number of false negatives.

### 4.2.1 Cross-validation

To verify the feasibility and stability of the constructed model, we performed cross-validation of the training data. We used a $k$-fold cross-validation scheme with $k$ corresponding to the split number of training and test data. For $k$-fold cross-validation, the dataset of each class is randomly divided into $k$ exclusive subsets to avoid grouping the data.

The principle states, for each value of $k$, to make $k$ iterations and at each iteration, $i$, $(k-1)$-fold of data are used for the training, then the $i$-th subset constructed for the test each time as we can see in Fig. 3. Thus, the final performance corresponds to the average performance of the $k$ iterations. We then calculated the performance of the model without cross-validation with $k$-splits (of which $k-1$ random subset for training and one for the test) and made the difference in absolute value between the value obtained and the performance value by cross-validation for SVM and DT models. We tested eight different values of $k$ chosen from 3 to 10 and illustrated the results obtained for accuracy and f1-measure in Fig. 4.

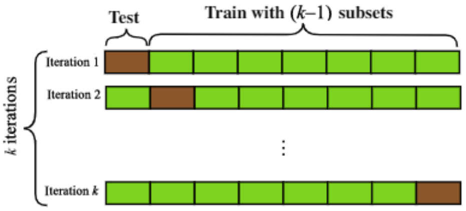


**Fig. 3 Principle of $k$-fold cross-validation.**

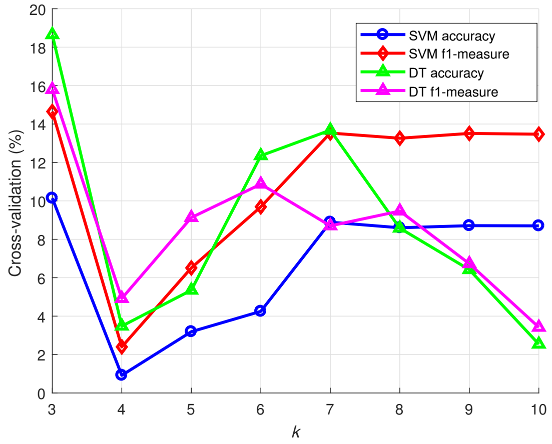


**Fig. 4 Variation of cross-validation according to the number of splits.**

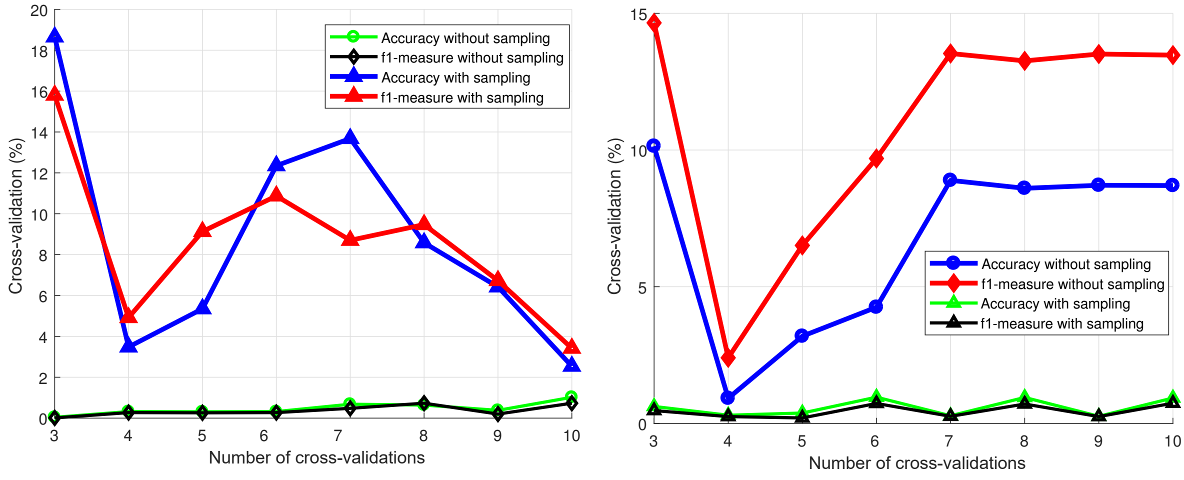
From Fig. 4, we notice that for different values of $k$, either accuracy or f1-measure, there is a considerable performance deviation for both SVM and DT algorithms. This difference is maximum for both accuracy and f1-measure for $k = 3$, where it reaches 10.5% and 15% for SVM and 19% and 16% for DT. On average, the difference is almost minimal for $k = 4$ for the two models (SVM and DT), where it varies between 1% and 5% and therefore remains significant. It shows that the data are not stable, and the model poses an overfitting problem due to the small size of the data and, therefore, it cannot be realistic. To overcome this problem and optimize the performance of the model, we applied the data resampling technique for each of the two models.

### 4.2.2 Resampling the data to improve performance

In general, resampling is necessary when the dataset has been recorded in a given time interval or includes a number of samples and we want to modify this interval or number of samples[2]. In this case, we want to resize and reorder the data in order to optimize the model's performances by eliminating the bias and the variance of the algorithms to build a stable and realistic model.

The analysis of performance deviation stabilization by cross-validation after resampling the data revealed that the compromise between the minimum of this deviation and the performance of the model is optimal for a data size equal to 300 of our samples. For this size of data, we illustrated in Fig. 5 that the deviation of $k$-fold cross-validation, for $k = 3, 4, \ldots, 10$, before and after resampling.

Figures 5a and 5b summarize the deviations of the $k$-fold cross-validation for the two performance metrics, namely accuracy and f1-measure according to the value



(a) Before resampling

(b) After resampling

**Fig. 5 Cross-validation before and after resampling of the soft tissues tumour data.**

of *k* before and after resampling the data. We can see that the deviation dropped sharply both for the accuracy and f1-measure after resampling the data. However, without resampling, the deviations in performance obtained by cross-validation for the accuracy and f1-measure are much higher. It confirms the importance of resampling data in order to optimize the performance of the model by reducing overfitting while establishing it for a more efficient and safer automatic prediction.

### 4.3 Global results analysis

In general, for computer-aided diagnostic systems, the most critical factor is their precision in terms of performance[2]. We have presented the performance using two different and more efficient models, SVM and DT. First, in Figs. 6 and 7, we illustrated the performances (accuracy, AUC, and f1-measure) of the model built based on DT and SVM algorithms, for the splits equal to 3, 4, then 5.

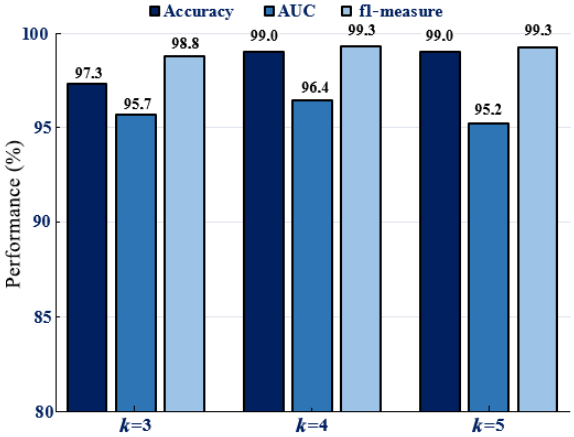


**Fig. 6 DT model performances.**

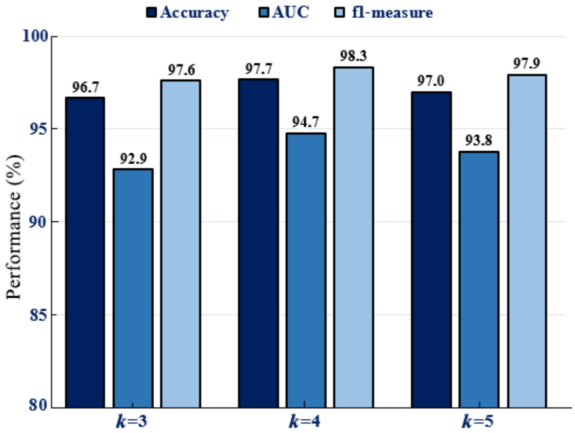


**Fig. 7 SVM model performances.**

According to Fig. 6, we found that overall for the three measures, the performances are better for four splits where they reach 99.0%, 96.4%, and 99.3% in terms of accuracy, AUC, and f1-measure, respectively. Then for five splits, we obtained for these three measures: 99.0%, 95.2%, and 99.3%, respectively; and finally, they are lower for three splits with 97.3%, 95.7%, and 98.8%, respectively for accuracy, AUC, and f1-measure.

As in the previous case (Fig. 7), we also found that the performances are better when using four splits with 97.7%, 94.7%, and 98.3% for the measurements of accuracy, AUC, and f1-measure, respectively. When using five splits, we obtained 97.0%, 93.8%, and 97.9% for these three measurements, respectively; and they are also still relatively weaker when using three splits with 96.7%, 92.9%, and 97.6% for accuracy, AUC, and f1-measure, respectively.

By making a comparative study of the results illustrated in Figs. 6 and 7, we noticed that the optimal number of data splits to have better performance and a stable model is four, both for SVM and DT algorithms.

Table 2 shows that for this number of splits, we noticed that the DT algorithm was slightly better compared to the SVM algorithm. The best performance of the DT-based model was 99.0% for accuracy, 96.4% for AUC, and 99.3% for f1-measure. Therefore, the DT-based model exceeded the SVM-based model by 2.0% in terms of accuracy, 2.6% in terms of AUC, and 1.4% in terms of f1-measure. Unlike the majority of classification problems, it has been found that the value of f1-measure is always higher than accuracy and AUC since the classes are unbalanced, where the instances classified as positive are twice of those classified as negative. This analysis led us to seek which proportions of these two models use the different features of the data to make the classification. In the next part, we will analyze the importance of features for these two models.

Regarding the importance of features and impact analysis (or their relevance) on the performance of each model, we can refer to Fig. 8 for DT and Fig. 9 for SVM.

From the classification illustrated in Fig. 8, we found that DT took into account two most relevant attributes, among which are the attribute "*ID*" with a score of importance around 27% and the relevance of the attribute "*Clotting-Time* (*CT*)" with a score around 18%. Apart from these two attributes, DT used just eight other attributes with importance scores varying approximately between 2% and 12%.

**Table 2 Best performance results of model-based on DT and SVM algorithms.**

(%)

| Algorithm | Accuracy | AUC | f1-measure |
|---|---|---|---|
| DT | 99.0 | 96.4 | 99.3 |
| SVM | 97.7 | 94.7 | 98.3 |

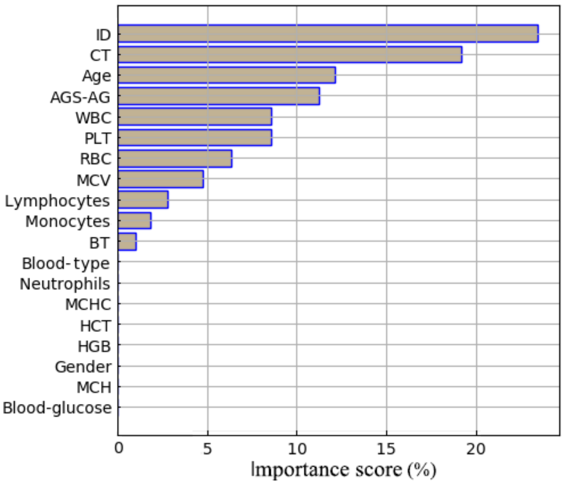


**Fig. 8 Importance of features for DT-based model.**

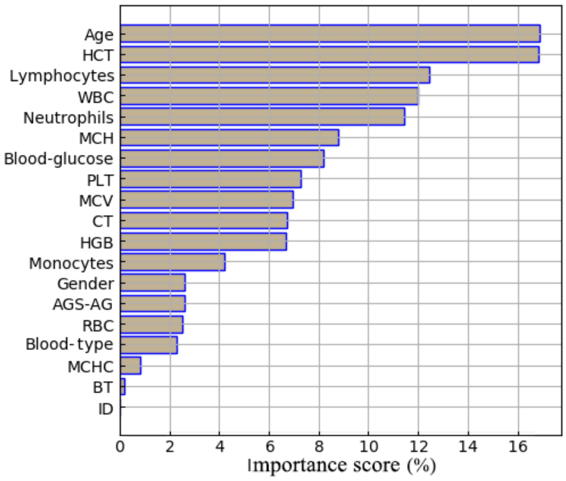


**Fig. 9 Importance of features for SVM-based model.**

According to the classification illustrated in Fig. 9, we noticed that SVM took into account all the attributes even if two attributes were relatively the same among which appear the attribute *age* with a score of importance around 17% and the relevance of the attribute *HCT* with a score also around 17%. Apart from these two attributes, SVM used all the other attributes with importance scores varying approximately between 2% and 13% except the attribute *ID* whose relevance for SVM is almost zero.

The most important feature for the model based on DT, and not necessary for the model based on SVM, is *ID*, which is not characteristic of a patient contrary to the attribute *age* but turns out to be the most relevant. Although the feature *gender* is specific for the profile of each patient, it was considered less relevant by the two models and appears so in Figs. 8 and 9. It does not necessarily mean that this type of attribute generally has little impact on modeling the success of separating STT and non-STT. In this particular case, the indicators for the blood tests used were measured by the hospital, and the results obtained suggest that these indicators are probably better suited to our problem of targeting soft dissolving tumors.

## 5 Discussion

In the previous section, we analyzed the performance of our model by comparing the DT and SVM algorithms. In this section, we compare these results with the performance of other algorithms and the previous work. Finally, we make a global discussion of the contribution of our work compared to the state-of-the-art automatic diagnosis.

### 5.1 Comparison with other ML algorithms

Table 3 recapitulates the comparative study of performance results of our two models (SVM and DT) with other ML algorithms, namely Logistic Regression (LR), $k$-Nearest Neighbor (KNN), Naive Bayes (NB), and Artificial Neural Networks (ANN). From Table 3, we can see that SVM and DT greatly exceed the performance of the other models for the three measures (accuracy, AUC, and f1-measure). In increasing order of performance, comes LR with 90.7%, 81.6%, and 94.1%, respectively for accuracy, AUC, and f1-measure; then KNN ($k = 3$ being the optimal $k$) with 88.0%, 63.2%, and 91.7%, followed by NB with 81.3%, 63.2%, and 88.9%. Finally, we also noticed that the model based on Artificial Neural Networks (ANN) gives the lowest performance, for the measurements used (76.0%, 52.6%, and 86.2%, respectively for accuracy, AUC, and f1-measure). The reason for this poor performance is due to the small data size.

### 5.2 Comparison with other previous works

Since the objective of an automatic classification model is to predict results as close to reality as possible, the main challenge of our study, we compared our results to the previous work carried out by Zahras et al.[34] on this same problem and summarized this comparison in Table 4.

In Table 4, we note that compared to the previous study of Zahras et al.[34], which also used the SVM and stochastic-SVM algorithms, there is massive improvement in the performance. In terms of accuracy and f1-measure, our model yielded an increase of 32.9% and 26.3%, respectively, above their SVM model. Their stochastic-SVM algorithm was more efficient than their SVM model having reached 57.82% and 66.00% for accuracy and f1-measure, respectively, in their study. Most recently, Rustam et al.[10] published additional work on this problem by improving the performance of Ref. [34] with Fuzzy C-means clustering and an SVM model, but the performance was still weaker compared to ours. Overall, we have improved the best current values (71.43% for accuracy and 83.33% for f1-measure)[10] on this database by 27.57% and 15.67 % in terms of accuracy and f1-measure, respectively.

**Table 3 Comparison with other algorithms and the corresponding performance of each of them. Here, MLP represents multi-layer perception.**

(%)

| Algorithm | Accuracy | AUC | f1-measure |
|---|---|---|---|
| DT | 99.0 | 96.4 | 99.3 |
| SVM | 97.7 | 94.7 | 98.3 |
| LR | 90.7 | 81.6 | 94.1 |
| KNN ($k = 3$) | 88.0 | 63.2 | 91.7 |
| NB | 81.3 | 63.2 | 88.9 |
| ANN (MLP) | 76.0 | 52.6 | 86.2 |

**Table 4 Comparison of the performance between our automatic classification model with previous work.**

(%)

| Algorithm | Accuracy | AUC | f1-measure |
|---|---|---|---|
| DT | 99.00 | 96.40 | 99.30 |
| SVM | 97.70 | 94.70 | 98.30 |
| SVM[34] | 57.82 | – | 66.00 |
| Stochastic-SVM[34] | 64.80 | – | 72.00 |
| Fuzzy C-means[10] | 71.43 | – | 81.82 |
| SVM[10] | 71.43 | – | 83.33 |

### 5.3 Overall discussion

Machine learning-based algorithms for automatic classification driven by data from analysis of patient medical records are potentially interesting tools for differentiating between STT and non-STT. We have demonstrated the detailed process of building a model that proves a near-perfect correlation of its predicted results with those of patient clinical records. Thus, a model could be an excellent computer-aided tool to assist doctors in the precise recognition of STT and non-STT.

Indeed, after evaluating the performance of our model based on SVM and DT algorithms in the previous parts, our main result was that the DT algorithm based model gave the best results for predictive classification of STT compared to the SVM algorithm. However, studying the classification attributes for both models allowed us to see that the SVM algorithm utilizes all the variables of the problem, particularly those which are characteristic of the patient's condition, as shown in Figs. 8 and 9. The DT algorithm, which uses only half of the problem variables, turns out to be very sensitive to overfitting and therefore presents a higher bias and variance compared to SVM algorithm, which turns out to be more realistic in terms of performance. We are also interested in the execution time of each of these algorithms. Although the SVM algorithm takes a relatively longer time than the DT algorithm, this time is still very low for both models. After this study, we are interested in a study comparing the other most commonly used algorithms to the methods used by radiologists. In general, radiologists can perform STT classification at around 90% for accuracy and 85% for AUC[2]. These numbers show that in addition to being a safe, fast, and stable model, our model could be more effective way of performing STT classification. To build our model, we used a dataset comprised of mainly patient blood analysis, unlike several other published works which most often used MRI analysis[2, 35–37]. The size of this dataset always represents a small random sample of the entire population of STT. Just like other researches[10, 34], we forced to resample these data in order to build the best model, but we detailed this resampling process supported by cross-validation of the results. We needed much larger training samples for practical applications to cover all the different pathological types of STT[2].

Many works have previously implemented classifiers to separate tumors, whether by data from analysis of MRI images, blood, or other types of data, but these models, often based on less than 50 samples, do not detail the different stages of model construction. Besides, they are most often limited to performance measures of non-exhaustive trained classifiers, which we tried to cover in this study. The comparison based on the nonparametric statistical test of the performance of DT and SVM algorithms with other classification algorithms showed that DT or SVM algorithms worked just as well or even better. Compared to the previous work done by the stochastic-SVM algorithm approach on this same dataset[34], we have a significant performance improvement of more than 32.9% and 26.3% in terms of accuracy and f1-measure, respectively. Moreover, we corrected several anomalies related to the process of the construction of classification models, which were recorded in this paper. In summary, the results obtained in our study were entirely predictable, knowing that the analysis of patient medical records by computer algorithms can quickly extract more information from tumor medical records compared to visual assessments done by radiologists. We can cite several examples: the analysis of the textures of MRI images in Ref. [2] and the analysis of the field of mammograms digitized in Ref. [38]. Therefore, we believe that a diagnosis derived from the analysis of medical records can play a similar role in the separation of STT from non-STT because it can subjectively (or even better) model that of a visual assessment by a physician.

## 6 Conclusion

High precision calculations enriched by ML algorithms can benefit applications in several disciplines, including the medical field. These tools have made it possible to improve the performance of computer-aided diagnostic systems significantly in recent years, and their integration continues to provide a challenge for modern health institutions. In this paper, we constructed a robust and realistic model, starting from data collected at the Nur Hidayah Hospital in Bantul, Yogyakarta, Indonesia, allows automatic predictive classification of STT and non-STT. After integrating a new data preprocessing technique, we compared two classifiers, namely SVM and DT algorithms. This comparison showed that even if the DT algorithm is slightly more efficient than the SVM algorithm, the DT model is much more sensitive to the number of variables than the SVM model. The implementation of a computer-aided diagnostic system based on our model could also prove to be more efficient and effective than the diagnosis carried out entirely by a visual assessment done by a radiologist and state-of-the-art models often analyzed from MRI images[2, 36]. In concrete terms, we have overcome and clarified the process of constructing the model, which has improved the best performance of the previous study carried out on this database[10, 34] of more than 27.57% and 15.67% in terms of accuracy and f1-measure, respectively. Our future research will focus on strengthening and continuously improving our model by combining weak algorithms in order to adapt it to the automatic diagnosis of other types of diseases such as glaucoma.

## References

[1] F. Collin, M. Gelly-Marty, M. B. N. Binh, and J. M. Coindre, Sarcomes des tissus mous: Donneés anatomopathologiques actuelles, *Cancer/Radiothérapie*, vol. 10, nos. 1&2, pp. 7–14, 2006.

[2] J. Juntu, A. M. De Schepper, P. Van Dyck, D. Van Dyck, J. Gielen, P. M. Parizel, and J. Sijbers, Classification of soft tissue tumors by machine learning algorithms, in *Soft Tissue Tumors*, F. Derbel, ed. London, UK: IntechOpen, 2011, pp. 53–69.

[3] A. M. De Schepper and J. L. Bloem, Soft tissue tumors: Grading, staging, and tissue-specific diagnosis, *Top. Magn. Reson. Imaging*, vol. 18, no. 6, pp. 431–444, 2007.

[4] T. Hayashi, A. Horiuchi, K. Sano, Y. Kanai, N. Yaegashi, H. Aburatani, and I. Konishi, Biological characterization of soft tissue sarcomas, *Annals of Translational Medicine*, vol. 22, no. 3, p. 368, 2015.

[5] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, Texture analysis of medical images, *Clin. Radiol.*, vol. 59, no. 12, pp. 1061–1069, 2004.

[6] Y. L. Huang, K. L. Wang, and D. R. Chen, Diagnosis of breast tumors with ultrasonic texture analysis using support

vector machines, *Neural Comput. Appl.*, vol. 15, no. 2, pp. 164–169, 2006.

[7] B. Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch, Inability of humans to discriminate between visual textures that agree in second-order statistics—Revisited, *Perception*, vol. 2, no. 4, pp. 391–405, 1973.

[8] H. Farhidzadeh, B. Chaudhury, M. Zhou, D. B. Goldgof, L. O. Hall, R. A. Gatenby, R. J. Gillies, and M. Raghavan, Prediction of treatment outcome in soft tissue sarcoma based on radiologically defined habitats, in *Proc. SPIE 9414*, *Medical Imaging 2015*: *Computer-Aided Diagnosis*, Orlando, FL, USA, 2015, p. 94141U.

[9] M. Karanian and J. M. Coindre, Quatrième édition de la classification OMS des tumeurs des tissus mous, *Ann. Pathol.*, vol. 35, no. 1, pp. 71–85, 2015.

[10] Z. Rustam, S. Hartini, T. Siswantining, D. A. Utami, and N. K. Putri, Comparison between fuzzy kernel C-means, fuzzy kernel possibilistic C-means and support vector machines in soft tissue tumor classification, in *Advanced Intelligent Systems for Sustainable Development* (*AI2SD'2019*), M. Ezziyyani, ed. Cham, Germany: Springer, 2020, pp. 92–105.

[11] H. S. Xu, L. Wang, and W. L. Gan, Application of improved decision tree method based on rough set in building smart medical analysis CRM system, *Int. J. Smart Home*, vol. 10, no. 1, pp. 251–266, 2016.

[12] P. D. Afonso and V. V. Mascarenhas, Imaging techniques for the diagnosis of soft tissue tumors, *Rep. Med. Imaging*, vol. 8, pp. 63–70, 2015.

[13] C. D. M. Fletcher, K. K. Unni, and F. Mertens, *Pathology and Genetics of Tumours of Soft Tissue and Bone*. Lyon, France: IARC Press, 2002.

[14] C. D. M. Fletcher, The evolving classification of soft tissue tumours: An update based on the new WHO classification, *Histopathology*, vol. 48, no. 1, pp. 3–12, 2006.

[15] C. D. M. Fletcher, The evolving classification of soft tissue tumours–An update based on the new 2013 WHO classification, *Histopathology*, vol. 64, no. 1, pp. 2–11, 2014.

[16] C. G. L. Guillou, Tumeurs des tissus mous: Rôle du pathologiste dans l'approche diagnostique, *Rev. Med. Suisse*, vol. 3, p. 32473, 2007.

[17] P. Marec-Bérard, F. Chotel, and L. Claude, PNET/Ewing tumours: Current treatments and future perspectives, *Bull. Cancer*, vol. 97, no. 6, pp. 707–713, 2010.

[18] K. Scotlandi, D. Remondini, G. Castellani, M. C. Manara, F. Nardi, L. Cantiani, M. Francesconi, M. Mercuri, A. M. Caccuri, M. Serra, et al., Overcoming resistance to conventional drugs in Ewing sarcoma and identification of molecular predictors of outcome, *Journal of Clinical Oncology*, vol. 27, no. 13, pp. 2209–2216, 2009.

[19] D. Komura and S. Ishikawa, Machine learning methods for histopathological image analysis, *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, 2018.

[20] C. S. T. Koumetio, W. Cherif, and S. Hassan, Optimizing the prediction of telemarketing target calls by a classification technique, in *Proc. 2018 $6^{th}$ Int. Conf. on Wireless Networks and Mobile Communications*, Marrakesh, Morocco, 2018, pp. 1–6.

[21] S. C. K. Tekouabou, W. Cherif, and H. Silkan, A data modeling approach for classification problems: application to bank telemarketing prediction, in *Proc. $2^{nd}$ Int. Conf. on Networking*, *Information Systems & Security*, Rabat, Morocco, 2019, pp. 1–7.

[22] K. Lakshminarayan, S. A. Harp, and T. Samad, Imputation of missing data in industrial databases, *Appl. Intell.*, vol. 11, no. 3, pp. 259–275, 1999.

[23] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, Decision tree and SVM-based data analytics for theft detection in smart grid, *IEEE Trans. Ind. Inform.*, vol. 12, no. 3, pp. 1005–1016, 2016.

[24] Y. W. Chang, C. J. Hsieh, K. W. Chang, M. Ringgaard, and C. J. Lin, Training and testing low-degree polynomial data mappings via linear SVM, *J. Mach. Learn. Res.*, vol. 11, pp. 1471–1490, 2010.

[25] N. A. Shrivastava, A. Khosravi, and B. K. Panigrahi, Prediction interval estimation of electricity prices using PSO-tuned support vector machines, *IEEE Trans. Ind. Inform.*, vol. 11, no. 2, pp. 322–331, 2015.

[26] S. S. Keerthi and C. J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.

[27] R. A. Lippert and R. M. Rifkin, Infinite-$\sigma$ limits for Tikhonov regularization, *J. Mach. Learn. Res.*, vol. 7, pp. 855–876, 2006.

[28] S. Ruggieri, Efficient C4.5 [classification algorithm], *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 2, pp. 438–444, 2002.

[29] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, An efficient rule-based classification of Diabetes using ID3, C4.5, & CART ensembles, in *Proc. 2014 $12^{th}$ Int. Conf. on Frontiers of Information Technology*, Islamabad, Pakistan, 2014, pp. 226–231.

[30] S. L. Salzberg, Book review: C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, inc., 1993, *Mach. Learn*, vol. 16, no. 3, pp. 235–240, 1994.

[31] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow*. Birmingham, UK: Packt Publishing, 2019.

[32] G. Marinakos and S. Daskalaki, Imbalanced customer classification for bank direct marketing, *J. Mark. Anal.*, vol. 5, no. 1, pp. 14–30, 2017.

[33] S. Young, C. H. Huang, and M. McDermott, Internationalization and competitive catch-up processes: Case study evidence on Chinese multinational enterprises, *Manage. Int. Rev.*, vol. 36, no. 4, 295–314, 1996.

[34] D. Zahras, Z. Rustam, and D. Sarwinda, Soft tissue tumor classification using stochastic support vector machine, *IOP Conf. Ser. Mater. Sci. Eng*, vol. 546, no. 5, p. 052089, 2019.

[35] Y. Zhang, Y. F. Zhu, X. M. Shi, J. Tao, J. J. Cui, Y. Dai, M. T. Zheng, and S. W. Wang, Soft tissue sarcomas: Preoperative

predictive histopathological grading based on radiomics of MRI, *Acad. Radiol.*, vol. 26, no 9, pp. 1262–1268, 2019.

[36] Y. Lee, J. B. Seo, J. G. Lee, S. S. Kim, N. Kim, and S. H. Kang, Performance testing of several classifiers for differentiating obstructive lung diseases based on texture analysis at high-resolution computerized tomography (HRCT), *Comput. Methods Programs Biomed.*, vol. 93, no. 2, pp. 206–215, 2009.

[37] J. Juntu, J. Sijbers, S. De Backer, J. Rajan, and D. van Dyck, Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images, *J. Magn. Reson. Imaging*, vol. 31, no 3, pp. 680–689, 2010.

[38] J. M. Boone, K. K. Lindfors, C. S. Beatty, and J. A. Seibert, A breast density index for digital mammograms based on radiologists' randing, *J. Digit. Imaging*, vol. 11, no. 3, p. 101, 1998.

**El Arbi Abdellaoui Alaoui** received the PhD degree in computer science from Faculty of Sciences and Technology, Errachidia, University of Moulay Ismaïl, Meknès, Morocco in 2017. Prior to this, he received the master degree in telecommunication from the National School of Applied Sciences, University of Sidi Mohamed Ben Abdallah, Fès, Morocco in 2013. He is currently a research professor at EIGSI, Casablanca and My Ismail University, Errachidia, Morocco. His research publications include mainly wireless networking, ad hoc networking, DTN networks, game theory, Internet of Things (IoT), smart cites, and optimisation.

**Stéphane Cédric Koumetio Tekouabou** received the PhD degree from the Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco in 2020. He obtained the MEng degree from the National School of Applied Sciences El Jadida (ENSAJ) of the same university in 2016. He has also participated in the scientific committee of many international conferences. His research interests include computer vision (recognition, detection, and classification problems), artificial intelligence, machine learning, deep learning, and optimization.

**Sri Hartini** received the bachelor and master degrees from Universitas Indonesia in 2019 and 2020, respectively. She is currently pursuing the PhD degree in intelligence computation at Universitas Indonesia. She is passionately carrying out researches on machine learning, computer vision, neural networks, and deep learning in various fields.

**Zuherman Rustam** is an associate professor and a lecture of the intelligence computation at the Department of Mathematics, Universitas Indonesia. He obtained the master degree in informatics from the Paris Diderot University, France in 1989, and completed the PhD degree from computer science, Universitas Indonesia in 2006. His research interests are machine learning, pattern recognition, neural network, and artificial intelligence.

**Hassan Silkan** received the PhD degree from Sidi Mohamed Ben Abdellah University, Fès, Morocco in 2009. He is a professor at the Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco. His research area is shape representation and description, similarity search, content based image retrieval, database indexing, 2D/3D shapes indexing and retrieval, and multimedia databases.

**Said Agoujil** received the PhD and MS degrees in mathematics from Faculty of Sciences and Technology of Marrakech (FSTM), Morocco in 2008 and 2004, respectively. He is currently a professor at the Department of Computer Science, Faculty of Sciences and Technology, My Ismail University, Errachidia, Morocco. His current research interests include numerical analysis, wireless network, linear algebra, and speech coding.