Machine Learning with Applications 13 (2023) 100480

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

# New interpretation of GNRB® knee arthrometer results for ACL injury diagnosis support using machine learning

Jean Mouchotte [a,*], Matthieu LeBerre [a], Théo Cojean [b], Henri Robert [c]

[a] Learnind Data & Robotics Laboratory (LDR), ESIEA, LAVAL, 53 000, France
[b] Univ Lyon, Univ Gustave Eiffel, Univ Claude Bernard Lyon 1, LBMC UMRT 9406, Lyon, F-69622, France
[c] Centre Hospitalier du Haut Anjou, Château-Gontier-Sur-Mayenne, 53 200, France

ARTICLE INFO

*Keywords:*
ACL tears
AI medical diagnosis
Arthrometry
GNRB®
Machine learning
Diagnostic accuracy
Knee anterior laxity

ABSTRACT

**Purpose:** GNRB is an arthrometer and alternative diagnostic method less expensive than MRI and more accurate than KT-1000 in Anterior Cruciate Ligament (ACL) tears detection. Dynamic knee laxity tests are more complex to analyze and will require a new solution of universal interpretation. The hypothesis is that using a solution based on Artificial Intelligence (AI) will allow us to obtain a more accurate and robust non-invasive diagnostic method than the current solution with three laxity thresholds.
**Method:** AI can enhance the reliability of this analysis by utilizing advanced algorithms and incorporating a wide range of additional parameters, leading to more precise diagnostics. The existing process solely rely on laxity differences obtained from the device, overlooking influential factors like clamping force. By considering a broader set of parameters and employing well-calibrated models a comparative study was performed between different Machine Learning (ML) models and Ensemble Learning to get the best compromise. The correction process will leverage statistical analysis of the current solutions.
**Results:** Association of Voting, Stacking and threshold laxity methods results report a 6% increase in accuracy and approximately 13% improvement in tear detection compared to the current solution with 1384 GNRB® measurements. Predicted diagnoses are also more prone to new data from patients unknown to the model and confirmed using a validation database.
**Conclusion:** A first ML model was introduced in ACL tears detection using GNRB device. GNRB coupled with ML was encouraging with better results than the current static diagnostic method. It could be integrated and recommended as a complementary solution to MRI.

## 1. Introduction

Anterior cruciate ligament (ACL) rupture is a common knee injury that involves complex movements such as cutting and pivoting. The diagnosis of this pathology is a public health issue because a patient must consult an average of three doctors to obtain a result (Micheo et al., 2010). The reliability of ACL tear diagnosis is currently dependent on the experience of the healthcare professional conducting the clinical examination, which includes Lachman test (Branch et al., 2010; Torg et al., 1976). However, in many cases, further examinations such as MRI may be required to confirm the diagnosis. MRI is the most commonly used non-invasive examination for diagnosing ACL tears, although partial rupture, scanning technique, or knee pain impacts its accuracy (Chang et al., 2013; Crawford et al., 2007; Ebrahimipour et al., 2014; Phelan et al., 2016). Less expensive and easier solutions without radiation exposure have therefore been developed. Moreover, it cannot directly assess knee laxity and is mainly used to assess the associated tears.

In the 1980s, Americans introduced the KT1000 with dynamic laximetry (Balasch et al., 1999). It measures the relative anterior tibial drawer movement to the femur, providing an objective measurement of tibial translation using an arthrometer. The Rolimeter (Schuster et al., 2004), developed by Roland Jacob, and stress radiography via Telos (Pässler & März, 1986) have been introduced as complementary diagnostic methods. However, many factors can affect the quality of the results, such as examiner experience, clamping force, positioning, or leg rotation. As a result, a large body of literature reports the inherent inaccuracy and poor reproducibility of these devices, resulting in subjective and unreliable results (Bouguennec et al., 2015; Collette et al., 2012; Jenny et al., 2017; Lefevre et al., 2014). Therefore, a new arthrometer, the GNRB®, has been developed.

* Corresponding author.
*E-mail addresses:* mouchotte@et.esiea.fr (J. Mouchotte), matthieu.leberre@esiea.fr (M. LeBerre), theo.cojean@etu.univ-lyon1.fr (T. Cojean), henri.robert36@gmail.com (H. Robert).

https://doi.org/10.1016/j.mlwa.2023.100480
Received 11 April 2023; Received in revised form 30 May 2023; Accepted 18 June 2023
Available online 22 June 2023

**Table 1**
Comparison of results obtained in 8 different studies using the GNRB device and/or MRI, based on two criteria: sensitivity (Sn) and specificity (Sp). Each study includes its own private data and selection criteria, resulting in a limited number of data points, which explains the heterogeneous results.

| | GNRB® | | | | MRI | | | |
|---|---|---|---|---|---|---|---|---|
| | Complete | | Partial | | Complete | | Partial | |
| | Sn | Sp | Sn | Sp | Sn | Sp | Sn | Sp |
| Robert et al. (2009) | 70% | 99% | 80% | 87% | | | | |
| Klouche et al. (2015) | 92% | 96% | 92% | 98% | | | | |
| Di Iorio et al. (2014) | | | 72% | | 84% | 92% | | |
| Beaurain (2020) | 73% | | 62% | | 76% | | 30% | |
| Lefevre N | 84% | 81% | 87% | 87% | | | | |
| Beldame J | 62% | 75% | | | | | | |
| Phelan et al. (2016) | | | | | 87% | 93% | | |
| Bohu et al. (2012) | | | | | | | 32% | |

The GNRB® device (Genourob, Laval, France) represents the pioneering automated dynamic laximetry (ADL) system. It is one of the most effective non-invasive methods with superior accuracy and reproducibility than other devices. Saravia et al. (2020). However, the diagnosis is solely based on the absolute difference in laxity measured at 134 N between the healthy knee and the pathological knee according to three thresholds suggest by the manufacturer (Genourob, Laval) : 0 to 1.5 mm (healthy) ; 1.5 to 3.0 mm : (partial tear) and above 3 mm (total tear)

This static approach is not universal and explains the large differences in sensitivity and specificity obtained in different studies that can reach more than 20% (Beaurain, 2020). Others question the relevance of the thresholds and ask that they be re-evaluated (Mouarbes et al., 2018).

The hypothesis is that relying solely on the difference in laxity measured at 134 N may not be sufficient to achieve reliable results for patients worldwide, considering the recommended thresholds. Therefore, a new method is proposed in this study, which using a Machine Learning (ML) model to incorporate a broader range of parameters such as age, gender, height, weight, tightness, compliance, and other relevant data to enhance the accuracy of the diagnosis. The following problem can thus arise : Would the set of measurements performed by the GNRB® and the patient's data allow us to obtain a more accurate and universal solution using Machine Learning? The study assembled the largest database with 1384 GNRB® measurements from healthy and injured patients to address this scientific lock.

Firstly, GNRB® limitations will be discussed to explain the contribution of the proposed solution to the field of medicine. For this purpose, complementary approaches based on Machine Learning will suggest, each with improvements to reach an optimal ratio of sensitivity and specificity. Finally, the best solution will be confirmed through cross-validation using a new database collected at the Brest Hospital (France) to affirm or refute the hypothesis.

Intelligent decision algorithms for ACL tear detection will, to our knowledge, attend a first in the literature. It could then allow obtaining a complementary diagnosis more reliable than the existing solutions thanks to the consideration of a large number of data and parameters. The decision to explore AI is based on its remarkable ability to "grasp general medical principles and apply them to new patients" (Obermeyer & Emanuel, 2016).

## 2. Materials and methods

Accurate clinical detection using non-invasive methods is crucial for determining the most effective treatment approach. Hence, this article primarily focuses on the GNRB® as an attempt to enhance the accuracy of its diagnosis. Nonetheless, arthroscopy is considered as the "gold standard" in numerous studies due to its high reliability, as it allows direct observation of the condition of the ACL by the surgeon (Crawford et al., 2007). Operative reports will serve as the basis for comparison throughout the study and provide the targets for ML models.

### 2.1. GNRB® and results limitations

The GNRB® is based on the Lachman (Torg et al., 1976) test but in an automated way to guarantee a high reproducibility and optimal knee flexion of 20° (Collette et al., 2012). While this explains the high reproducibility and accuracy, the universality of GNRB® diagnostics can be questioned. Table 1 shows a significant heterogeneity of results which can be explained by :

- an apparent lack of transparency for use especially for the clamping force which impacts the Laxity measurement (Alqahtani et al., 2018; Bouguennec et al., 2015). The GNRB® documentation recommends a maximum difference of 20 N between two measurements;
- the fact that the difference in laxity measured at 134 N is not universal and uses their own thresholds. Studies suggest a re-evaluation, especially for the 1.5 mm threshold (Mouarbes et al., 2018);
- the GNRB® allows for multiple measurements from 67 N to 200 N (Robert et al., 2009), but not all of the data are exploited. Other factors could also be considered, including patient gender, which has not been reported;

Due to more consistent and accurate results (Table 1), MRI remains the best method to date. On the other hand, dynamic laximetry, which allows economic and societal benefits while avoiding radiation, needs to be improved to be recommended in the same way as MRI (Gustafsson et al., 2020).

### 2.2. DataBase design

The first step is to obtain a balanced and realistic database because all the models proposed afterward will depend on the choices made.

#### 2.2.1. GNRB® data selection

A GNRB® database of 30,000 measurements performed on 1840 patients over a period of 2 ± 7 months after injury between 2008 and 2019 with an average of 30 ± 12 years old. MRI scans are also and have been performed in several centers using a Siemens or Philips 1.5 T MRI scanner according to a standard protocol for the knee: sagittal fat-suppressed (FS) T2-weighted, sagittal proton density (PD)-weighted, axial FS PD-weighted, coronal T2 and PD-weighted images. No sagittal or coronal oblique planes were performed in this series.

They have allowed us to prepare a batch of 5000 healthy patients and 869 with arthroscopically confirmed ACL rupture. Each consists of a unique set of measurements at diverse forces (between 67 N and 250 N) on one healthy and injured knee (or two healthy knees) of the same patient. They had to be made the same day under similar conditions (clamping force less than 20N). All the choice explains why the number of exploitable data is voluntarily reduced.

The training and test base contains 85% of patients against 33% in the ideal because there are three classes. A model with the dominance

of one group in its training set will tend to predict the majority class and must be corrected. An arbitrary selection of healthy patients reduced them to 515 (36%) with 492 complete tears (36%) and 377 partial tears (28%). Partial LCA ruptures represent less than 30% because they are less operated than complete ruptures.

#### 2.2.2. Input parameter selection

Machine learning algorithms require that the input parameters be carefully chosen and present on all samples. Parameters allow the models to establish links between the training data and the arthroscopy diagnoses that are the reference. Above 150 N, a modest proportion of patients have laxity measurements. To maintain the 1384 samples, we could keep GNRB® tests performed only at 67 N, 89 N, and 134 N. The same problem is observed in Klouche's study (Klouche et al., 2015) on the difficulty of reproduction in the concrete case if the force applied is too great. The slope coefficient P2 calculated using the laxity measurements is retained. This aligns with Bercovy's (Bercovy & Weber, 1995) description that laxity alone is not a sufficient parameter to describe the biomechanical behavior of the ACL

Clamping force is accepted as a critical parameter (Alqahtani et al., 2018; Bouguennec et al., 2015) and can be a new input parameter in the same way as the patient's gender or age. Pearson's coefficient confirms that the clamping force impacts the quality of the measurements with a score above 0.40 (one knee) and 0.22 (laxity difference). Gender had similar scores, unlike age (score $< 0.10$), which was therefore ignored as an input parameter. These results are consistent with a recent study (Klasan et al., 2020) showing that male investigators had a significantly higher mm laxity reading than females.

The input parameters are the laxity measurements at three translation forces (67 N, 89 N, and 134 N), the P2 coefficient, the clamping force, and gender.

#### 2.2.3. Validation of the panel with respect to GNRB recommendation

Once the complete database has been created, it seems important to check the consistency of the data. The GNRB® documentation advises to keep the measurements only if the difference in clamping force is less than 20 N. Retain 5% ineligible data to add a bias to the models by imposing a greater generalization on them (Fig. 2).

Fig. 2 shows that the data are consistent given the three thresholds analysis with a specificity above 95%. However, several samples with partial and total tears have a laxity difference of less than 1.5 mm confirming the previous state of the art. The accuracy observed with this dataset is 70% with the three threshold diagnosis methods and is therefore consistent. The current solution does not seem robust to a random selection of measurements, although it is admissible.

### 2.3. Data mining using machine learning

The second stage is to establish the Machine Learning algorithms to maximize the differentiation between a healthy and an ACL tear.

#### 2.3.1. Most popular techniques

No type of machine learning algorithm is better than the others. The choice depends on the problem, the number of input variables (6), samples (1384), and the issue. Our problem is to predict a diagnosis among three classes: healthy (0), partial tear (1), or total tear (2). Algorithms like Naïve Bayes used for textual prediction are not suitable.

Regression, Clustering, and neural network techniques (Mahesh, 2020; Zhou, 2021) have been implemented with the database. The first two methods are unsuitable for the problems because there is no linear relationship and different scales between the data. In the case of neural networks, interpretability was complex, and the limited amount of data made it difficult to ensure reliable and robust results. Increasing the amount of various data would allow us to solve this problem, but we decided not to study it because of the risk of artificial data (Hairy, 2021). All these techniques have therefore been discarded.

Support Vector Machine (SVM) (Zhou, 2021) techniques reduce the classification problem to a hyperplane. Several models were trained (e.g SVC — Support Vector Machine or BGD — Batch Gradient Descent), but only SGD was retained due its high sensitivity. The results using the SVC algorithm seems promising because the separation is likely to be a straight line parallel to the x-axis around 1.5 mm. The choice was to select SGD due to its better performance (speed & accuracy) and ability to adapt to new data in real-time. It is the first study with AI to diagnose ACL tears, and new data can be added regularly, which is a significant advantage.

The 3D representation (Fig. 2) shows that each patient has at least four neighbors with the same result. The K-plus-neighbor algorithm is based on this principle and makes a prediction based on the similarities of a reference set. In addition to being easy to implement and highly adaptable to complex data, Knn is robust to noisy data. It should be accurate, especially considering that patients are in batches of 4-5 with the same diagnosis. Ultimately, it aims to approach human analysis by comparing measurements with those of other patients who have similar results to make a diagnosis. Consequently, this model should always be included as a reference due to its qualities and adaptability to the problem.

Decision trees are based on the same principle as GNRB®'s analysis with higher conditions and could be worthwhile to exploit. The decision trees were excluded due to the requirement for a shallow depth to avoid overfitting. As a result, the diagnoses were primarily based on differences in laxity, leading to inferior results. A single tree cannot capture the relationships between all the parameters. It excludes variables considered less significant than, for example, laxity. An alternative approach would be to use a random forest instead of a single tree, which provides a robust solution that incorporates all variables.

#### 2.3.2. Ensemble learning

Ensemble learning techniques are based on the assumption that combining methods with different predictions would result in a new, more accurate solution (Learnia, 2022; Zhou, 2021). However, three conditions had to be fulfilled :

1. no model should have an accuracy of less than 50;
2. the predictions of each model must be sufficiently different;
3. a large amount of data is required;

The Bagging consists (Learnia, 2022; Zhou, 2021) of training an algorithm on different parts of the training set to obtain the most varied predictions possible. Of all the labels, the one most present will correspond to the predicted diagnosis. One of the most popular algorithms is RandomForest (RF), which should enable effective detection of ACL ruptures despite sometimes very close measurements. It combines predictions from multiple decision trees, which helps reduce the bias and variance inherent in a single decision tree. This allows for a robust interpretation of the data, even in cases where the relationships are non-linear and the interactions between variables are subtle. Thus, this model addresses the issues associated with previous decision trees, making it an excellent choice.

Boosting algorithms (Learnia, 2022; Zhou, 2021) such as GradBoost and AdaBoost were excluded due to poor sensitivity–specificity trade-off. The main criterion for their exclusion is a high sensitivity to outliers as well as the risk of overfitting. Database include measurements that are heterogeneous depending on the clamping force and the patient, who may be hypermobile. Although these flaws can be mitigated, early implementations were strengthened because a decrease in diagnostic accuracy was noticeable.

The machine learning algorithms of Voting and Stacking have been subject to a more in-depth study due to their numerous advantages. The first advantage is that combining multiple individual models should result in a final model that is more accurate, robust, and generalizable. The second advantage is the ability to model more complex phenomena than what an individual model could achieve. Among the most accurate models mentioned in the previous section, a wise choice has to be made:

**Table 2**
Comparison of several supervised learning algorithms and thresholds methods using accuracy, sensitivity, and specificity based on three diagnoses: healthy, partial tear, or complete tear. The incorporation of tear-type precision highlights the most accurate method for classifying the type of tear. Conducted using cross-validation and 1384 GNRB measurements.

| | Thresholds | Knn | SGD | RForest | Vote | Stack | New_Method |
|---|---|---|---|---|---|---|---|
| Accuracy (all diagnosis) | 69.49% | 69.56% | 67.80% | 68.62% | 69.72% | 69.35% | 75.64% |
| Sensitivity (all diagnosis) | 80.71% | 91.50% | 84.26% | 92.70% | 91.80% | 92.77% | 93.32% |
| Specificity (all diagnosis) | 97.30% | 91.55% | 96.28% | 89.18% | 91.52% | 90.01% | 94.50% |
| Precision for partial ACLs | 40.95% | 38.62% | 15.48% | 38.63% | 37.89% | 36.32% | 56.02% |
| Precision for complete ACLs | 62.40% | 70.24% | 77.97% | 60.92% | 71.58% | 73.00% | 70.67% |
| NPV | 74.79% | 86.39% | / | / | 86.81% | / | 89.06% |
| PPV | 98.07% | 94.84% | / | / | 94.84% | / | 96.72% |
| PPV_Partial | 56.30% | 52.16% | / | / | 52.39% | / | 62.82% |
| PPV_Complete | 69.66% | 61.88% | / | / | 61.77% | / | 69.05% |

- voting method (Learnia, 2022; Zhou, 2021) use different algorithms trained on the same data. The difficulty lies in their selection and the weighting of their predictions for the final decision. The main advantage of this technique is the identification of the patterns that have contributed to the final predictions. Several configurations showed that it was essential to keep the KNN model (with weight = 2). The choice was to associate only Forest model (weight = 1) to emphasize lesion detection (specificity). No other models were added, such as SGD, as it did not improve accuracy. The best configuration combines 2KNN + 1RForest to achieve optimal distinction between healthy and affected patients.
- stacking method (Learnia, 2022; Zhou, 2021) replaces the majority label selection of the voting technique by applying a model named "MetaClassifier" (e.g. LogisticRegression). It will come to play the role of the judge to decide which prediction has the best chance to be valid. This time, a third model will be added to those selected during Voting (Knn & RForest). The objective is to add predictions that are significantly different in order to force the meta-classifier to generalize. The choice has been made for SGD in order to maximize the detection of total admitted ruptures. The low precision for classifying an LCA as healthy should enhance sensitivity due to the high precision of Knn & RForest. This is a compromise as sensitivity is expected to decrease. Therefore, the SGD, KNN, and RandomForest configuration may not be optimal. However, it is the best configuration obtained after successive evaluations involving multiple models;

### 2.4. New diagnosis process

The study could have stopped at the evaluation of the last two models, Voting and Stacking, as it already represents an innovation in the absence of studies for diagnosing an ACL rupture. The hypothesis was made that it was possible to find new interpretations of the data and machine learning models to achieve even more accurate diagnoses. This section aims to innovate by combining the current static methods with an AI model, for example. The hypothesis is that using the most accurate model (Voting) and reinforcing its decisions with a precise rupture detection model (Stacking) and another model for healthy patients (Three thresholds) would be relevant (Table 2). Additionally, new interpretations of the data can be made.

#### 2.4.1. Input parameters and interpretation

The first modification was to increase the impact of the laxity measurement to 134 N based on Klouche's assumption that a higher translation force will result in higher accuracy. The experiments showed that there was no gain but a decrease in accuracy. The fact of not observing a gain remains coherent because it allows a greater diversity of measurements and a generalization of the model thanks to a dynamic analysis of the data.

A second assumption is that a healthy patient would obtain a measurement close to zero against negative values in an ACL tear. Changing the interpretation of laxity could increases accuracy and improves the discrimination between healthy. Consequently, the newly implemented diagnostic method will use the relative difference. This is a first key differentiating factor in implementing a new method in addition to AI and should increase sensitivity with limited impact on specificity to achieve a rate beyond 90%.

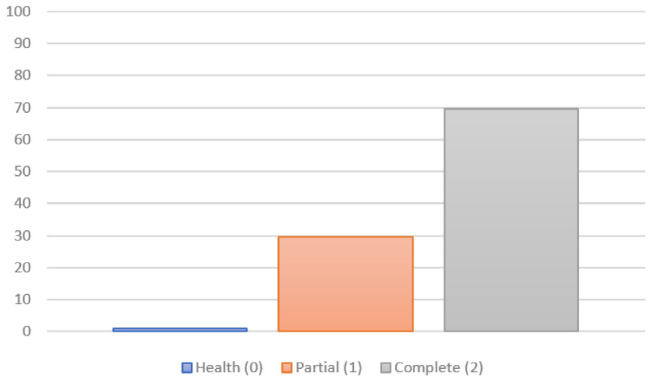



**Fig. 1.** The results of the statistical analysis demonstrate the probability that a diagnosis based on the proposed 3 mm threshold is correct. This threshold is recommended by the GNRB device manufacturers.

#### 2.4.2. Increasing sensitivity and ACL tears distinctions

This part proposes a solution that no longer involves comparing but rather associating three diagnostic methods (Voting, Stacking, Thresholds) in order to leverage the advantages of each method. Indeed, this approach involves creating a meta-model, similar to ensemble learning methods like voting, by combining multiple diagnostic methods together.

First change is to support or modify voting decisions based on statistical analysis. The study analyzes the distribution of measurements according to the diagnoses to demonstrate the proportion in which a diagnosis is correct based on the thresholds suggested by the GNRB device manufacturer (Fig. 1). It allows for influencing the predicted result by the voting method using the thresholds in two well-defined cases:

- beyond this threshold, there are less than 1% of patients without tears. A complete tear will then replace the AI prediction in the case of a healthy patient;
- there are 30% of partials and 70% of complete tears beyond 3 mm. If the result is a tear (partial or total or $\Delta > 3.0$ mm), a probability table ([0; 0.30; 0.70]) will be added to the predictions to accentuate the diagnosis in favor of a total rupture;

Beyond the change in the interpretation of values, which is a new approach, reinforcing the diagnosis through statistical analysis is a significant addition. This modification affects the probability of a diagnosis being true after a prediction by the trained models. This addition is a deliberate choice to improve the distinction between partial and complete ruptures. The final modification involves reusing part of this

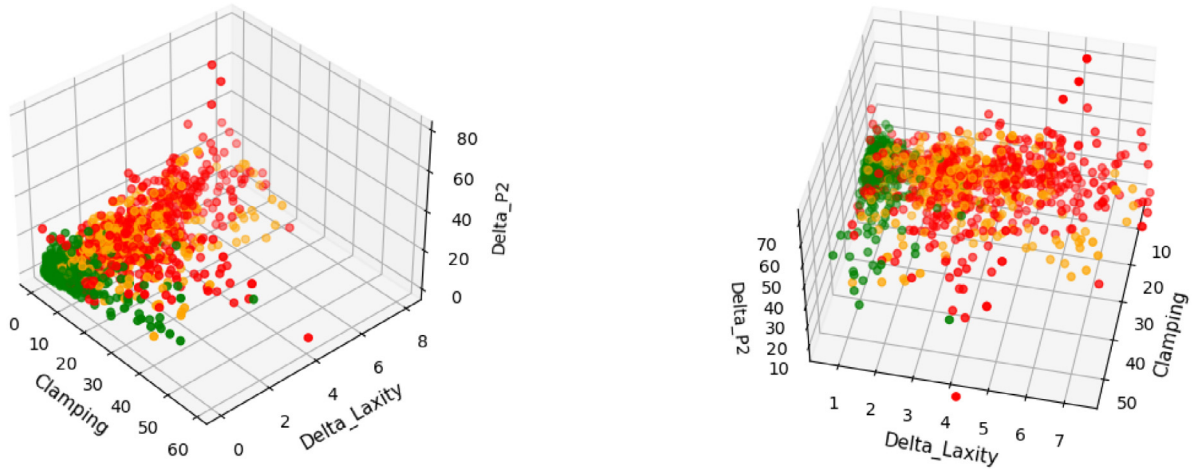



**Fig. 2.** Laxity measured of healthy and torn patients at 134 N as a function of clamp force, P2 slope coefficient with an arthroscopy result : healthy (green) ; partial tear (orange) ; complete tear (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analysis to associate, rather than compare, the solutions (Voting and Thresholds) by incorporating the Stacking model as a decision criterion. The aim is to achieve a model with the highest possible accuracy.

The second modification involves using the stacking and thresholding solution to improve the diagnostics made by the Voting method. The stacking ensemble learning method has proven to be the most accurate technique for detecting and classifying a rupture as complete (Table 2). Therefore, its combination with the thresholds defined by the manufacturers should increase the sensitivity and qualification of a rupture in two cases:

- to assert a complete rupture if the difference in laxity is superior at 3 mm by ignoring that of the voting method
- to correct a diagnosis of a healthy patient if the difference in laxity is greater than 1.5 mm and the Stack model admits a rupture. The diagnosis will be a partial tear as the Voting method defines the patient as healthy.
- in addition, in the case of a lesion detected by the Voting model, it is also possible to increase the rate of partial lesions. If the laxity difference is less than 1.5 mm and the Stacking model predicts at most partial rupture, the diagnosis will be a partial tear. Otherwise, the prediction of the Voting model will be retained as it is the most accurate, especially if the laxity difference is between 1.5 mm and 3.0 mm.

### 2.4.3. New diagnostic process with diagram explanation

By incorporating all the suggested improvements from the previous subsections, it is possible to achieve a new diagnostic process. This process combines statistical analysis to reinforce the decisions of the selected ensemble Machine Learning models: Voting (2Knn + RForest) and Stacking (Knn + SGD + RForest). The aim is to achieve a better trade-off between sensitivity and specificity while improving the qualification of a rupture as either partial or complete. Fig. 3 represents the new diagnostic suggested as a process diagram.

Firstly, the left part aims to improve the qualification of ruptures defined as partial or complete. Once ACL tear is diagnosed by the Voting method, a first condition ($\Delta < 3.0$ mm) aims to align with the analyses conducted in Section 2.4.2:

- When exceeding 3 mm and the second condition (stack_label=2) fulfilled, it will be possible to classify the rupture as complete, regardless of the label predicted by the first model (Voting). Otherwise, the probability table is added to the one provided by the voting decision. The aim is to reinforce the idea that a patient has a complete rupture without forcing it.
- On the contrary, if the laxity difference is less than 3 mm and 1.5 mm and the second model (Stacking) does not detect a complete rupture, it will be reclassified as partial. The hypothesis is that considering a total tear is likely an overstatement if two methods consider the LCA healthy or partially tear. If the condition ($\Delta < 1.5$ mm + stack_label $< 2$) is not met, the most accurate method (voting) is used.

The right part of the diagram aims to increase sensitivity and the rate of correctly diagnosed partial ruptures :

- The first condition ($\Delta < 1.5$ mm) confirms the prediction made by the Voting model since this diagnostic threshold accurately identifies patients as healthy.
- The second condition (stack_label = 2) reclassifies the initial diagnosis (vote_label=0) as having a partial lesion if the Stacking model predicts a tear. This modification should improve the diagnostic rate as Stacking is more precise in detecting a rupture, and the 1.5 mm threshold supports its choice.
- The last condition follows the analyses in Section 2.4.2, showing that ACL tears should qualify when the absolute laxity difference is above 3 mm. To support a diagnosis in favor of a rupture (partial or complete) without imposing the decision probability table will be applied.

### 2.5. Statistical analysis

Statistical analyses were performed using XLSTAT (Addinsoft, Paris, France), a software suite for data analysis and statistics in Microsoft Excel (Microsoft Corporation, Redmond, Washington, US). Confusion matrices were used to compare the ML models and assess the significance of differences, using sensitivity, specificity, positive predictive value, and negative predictive value as indicators of proportions. Significance was set at $p < 0.05$ for all analyses.

Simultaneously, the Matplotlib visualization library plot ROC curves for five different approaches evaluated on the same portion of the dataset (training and testing). It provides an overall measure of the performance of a machine-learning model and facilitates the comparison of different approaches.

All hyperparameters were automatically selected to achieve the best trade-off between accuracy and loss during model training. Choices were checked manually to ensure consistency, such as the optimal number of neighbors K, which was 6, consistent with an interpretation of Fig. 2.

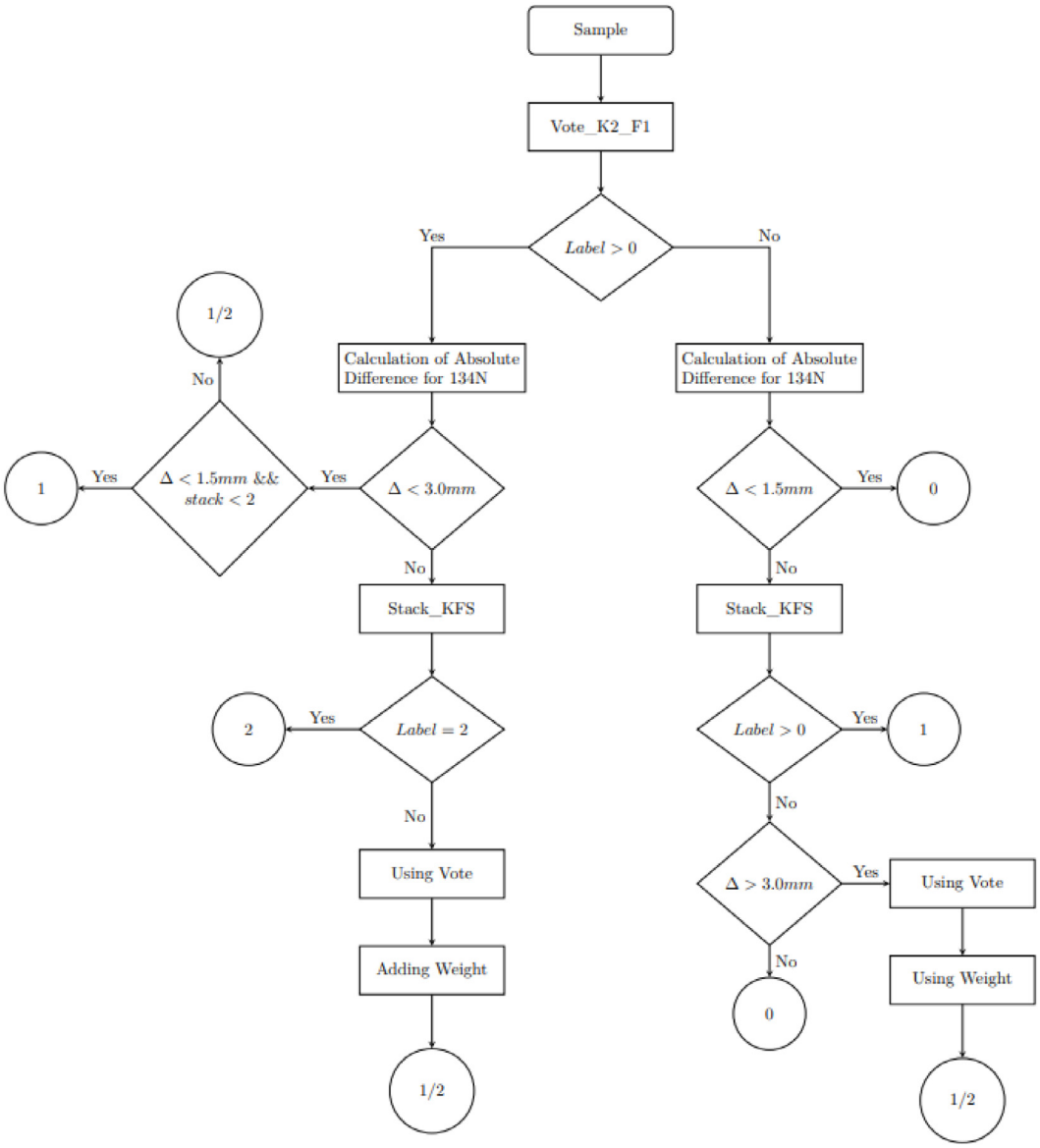



**Fig. 3.** Visualizing the Process of the Proposed New Diagnostic Methods through Graphic Representation.

## 3. Results

All methods compare the same data using successive cross-validation with the same parameters: sex, relative difference of laxity (67 N, 89 N, 134 N) and P2 coefficient. The objective is to ensure that the predicted diagnoses are not arbitrary or too specific to the training data. They use the relative difference in laxity measurement because it increases the accuracy by 4% and the distinction between my two types of ACls tears by 2%.

### 3.1. AI method comparisons

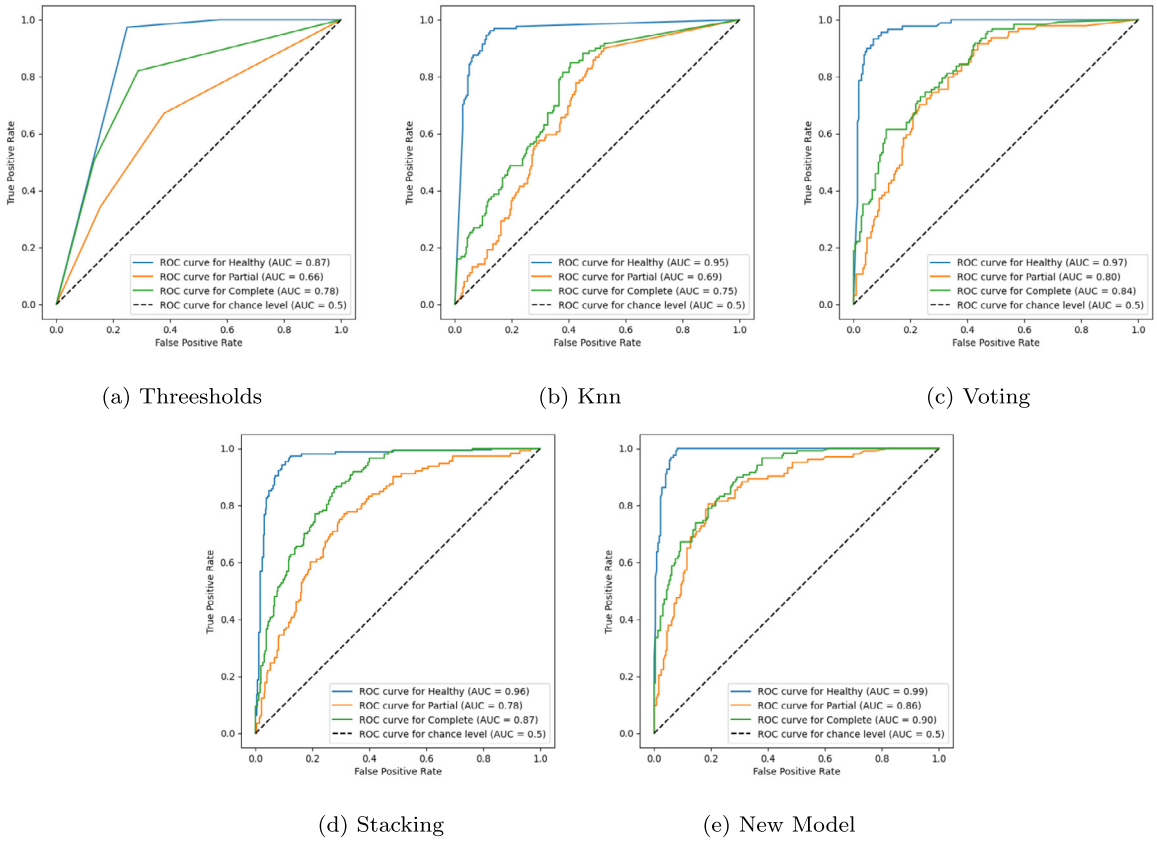
The accuracy of decision trees was about 5% lower than that of SGDs, leading to their exclusion. SGD was the second-best method, with a sensitivity close to 90%, excluding ensemble learning algorithms (Table 2). By using standard normalization to maximize accuracy, the KNN algorithm achieved a detection rate (specificity) of over 85% and sensitivity above 90%.

Based on the experimentation results presented in Table 2, the GNRB® solution recommended for diagnosis has an accuracy equivalent to 70% as three other methods. However, artificial intelligence algorithms outperformed the three-threshold solution, achieving a diagnosed rupture rate of 81% and an accuracy of 69.50%. Although the Random Forest algorithm did not perform as well as KNN, it could improve specificity. This explains why the Vote method, which combines multiple algorithms, can correctly distinguish a healthy ACL from a ruptured ACL with rates higher than 91% (sensitivity + specificity) and appears to be the most efficient solution at this stage.

The ROC curves demonstrate (Fig. 4) that the solution using the three thresholds is significantly less effective than the AI solutions, as it consistently falls below the other models. The curves for the Voting model consistently outperform the others, indicating that this model has better classification performance across all decision thresholds. However, Table 2 shows that the Stacking model has an area of 0.87 compared to 0.84 for complete injuries, confirming its usefulness in qualifying this type of rupture. The standalone KNN model performs worse than the Voting and Stacking models, further reinforcing the choice of model combination to achieve even higher precision.

### 3.2. Evaluation of the proposed new AI-based method

The new ML model provided results that were at least slightly inferior or superior to the thresholds method with no significant differences




(a) Threesholds (b) Knn (c) Voting

(d) Stacking (e) New Model

**Fig. 4.** Comparative evaluation of machine learning models (KNN, Voting, Stacking) and the reference method using the three laxity thresholds to assess the improvement of the new model through the analysis of the area under the curve (AUC).

($p > 0.244$) for Specificity, VPP, and VPP_Complete. For other results like Accuracy, Sensitivity, Precision (partial or complete), VPN, and VPP_Partial, they were better with a significant difference ($p < 0.001$). It explains significant gain for qualifying a tear as partial is greater than 56% (Table 2) with no drop in the rate for total ruptures. Table 2 also confirms that for positive and negative predictive values (PPV and NPV), our method is the most efficient with an optimal error rate. Finally, the best discrimination and accuracy between a healthy and a torn ACL is obtained with rates above 93%, justifying the choice to validate all the decisions and hypotheses of part 2.4.

Comparing the new ML method with other individual ML models, results were also at least equivalent with no significant differences ($p > 0.189$) for Sensitivity except for SGD with better results ($p < 0.001$), Specificity, Precision for complete ACLs except for SGD where results were worse ($p < 0.001$), and for RForest where results were better ($p < 0.001$), VPN and VPP. Results were significantly better ($p < 0.001$) compared to the other ML models for Accuracy, Precision for partial ACLs, VPP_Partial, and VPP_Complete.

When comparing the areas under the curves (Fig. 4), the proposed solution consistently outperforms the others, although the difference remains small compared to the Voting model, which is the second most accurate method. The new model is the only one to achieve an area of 0.90 for detecting complete rupture and almost 1.0 for detecting partial lesions (Fig. 4). These results demonstrate its capability to accurately classify all three possible diagnoses (healthy, partial tear, or complete tear). The gains are around 0.05 compared to traditional ML solutions and nearly three times higher with the static analysis based on laxity difference using three thresholds (Fig. 4).

MRI is recommended to establish an anterior cruciate ligament tear, whether partial or complete. However, only 169 patients have an MRI result and a valid GNRB test. Each measure is removed from the training base to compare their diagnoses with MRI diagnoses. The reference remains the arthroscopy result.

The comparison shows that the proposed solution detects ruptures better than MRI (sensitivity) by more than 1% (Table 3). The difference is insignificant since some of the MRI results are from the 2010s. Furthermore, MRI provides a better diagnosis if the rupture is total. On the other hand, GNRB® with Machine Learning remains more efficient in differentiating a partial tear from a complete tear (Table 3). Significant gain proves that this device is interesting to use before, after, or without MRI because it is less expensive and simpler to implement. Only the threshold method, which reaches barely 70% detection, is inferior and should not be recommended.

### 3.3. Validation of the robustness

The Brest Hospital could evaluate the proposed model using 88 GNRB® measurements from former patients following the device's usage recommendations. This is the first study that utilizes a second dataset established from a different source to validate the quality and robustness of its solution. The practitioners and the device used in this study are entirely independent from the training dataset, which consists of 1384 measurements.

The thresholds allow for a rupture in 87.80% compared to 92.68% with our solution. The difference in sensitivity with the results in Table 2 is 3%, compared to over 20% with the three thresholds (Table 1). Results demonstrate the robustness and quality of the proposed AI-based solution even in the face of new data. Finally, the MRI results showed that the rupture was not qualified for 2 of the 88 patients in this ultimate study. The same observation is made with our solution but with two different patients and allows us to conclude that the two solutions are at least equivalent to ACL tears detection.

**Table 3**
Results obtained to detect and qualify the type of tears in 169 patients with a GNRB and MRI result linked to an operative report. Specificity does not exist, as no patients are included in this database. The incorporation of the accuracy of the type of tear highlights the most accurate method between MRI, the three alone or the proposed solution based on ensemble learning to classify the type of tear.

| | Thresholds | New_Method | MRI |
|---|---|---|---|
| Accuracy (all diagnosis) | 37.56% | 62.64% | 54.91% |
| Sensitivity (all diagnosis) | 68.84% | 93.21% | 92.92% |
| Specificity (all diagnosis) | / | / | / |
| Precision for partial ACLs | 39.79% | 50.27% | 28.38% |
| Precision for complete ACLs | 35.87% | 71.83% | 74.95% |

## 4. Discussion

The most important finding of this study is that using three thresholds is not recommended despite previous study results. Indeed, our diagnostic proposal is better than that of Beaurin F and Beldame J (Beaurain, 2020) . The absence of pre-selection of data in favor of a diagnosis and the significant increase in data demonstrate the limitations of this approach according to the Table 2. The study confirms previous observations, particularly regarding the threshold of 1.5 mm, which requires reevaluation (Klouche et al., 2015; Mouarbes et al., 2018). Machine Learning applications combined with laxity measurement statistics allow for a much more precise approach (+5%). The main strength of the created method is the sensitivity and specificity ratio of around 94%, with a detection rate of partial ACL tears beyond 55% (Table 2). The PPV (positive predictive value) and NPV (negative predictive value) demonstrate that relying solely on laxity is unsufficient to ensure a reliable diagnosis in a patient.

Considering additional parameters like sex, grip strength, or laxity at different translating force demonstrated that GNRB® offers a more accurate solution than MRI (Table 3). An explanation is the better differentiation between partial and complete ACL tears (+10%) and improved detection of ruptures (+0.3%). Results support the findings of a 2011 (Van Dyck et al., 2011) study that showed that a standard protocol 1.5-T MRI performed by experienced radiologists might not be sufficiently reliable for diagnosing anterior cruciate ligament rupture. Instead of, comparison with a more recent study demonstrates that the results are similar to MRI diagnoses (Zhao et al., 2020). The accuracy is 92% for our proposed solution compared to 95% for MRI, favoring the latter. Detection of ALC tears (partial and total) shows 95% for MRI and 93% for GNRB (Table 2). The results for distinguishing between the two types of ruptures are very high compared to those reported in the literature (Table 1). This difference can be attributed to more recent data and the smaller number of data points available, with only 66 cases having an MRI compared to 256 (169/1384 + 88 from Brest) in this study. The study (Zhao et al., 2020) had to exclude patients due to strict selection criteria, such as a history of knee pathologies. Also, some people cannot do IRM as pregnant women or individuals with medical devices that prevent MRI. Further supports the value of offering an alternative diagnostic solution that is equally accurate, more cost-effective, and avoids radiation exposure.

The study does not suggest that GNRB® is superior to MRI because it is essential in clinical practice and cannot be replaced. Furthermore, the use of conclusions from MRI reports is subject to interpretation. They may vary depending on factors such as the level of experience of the radiologist, which could explain the observed differences in results.

The most controversial point is the difference in results between this study and existing studies. In our case, sensitivity and specificity with multiple datasets (test base, MRI base, validation base) vary by a maximum of 3%, which is low compared with 20% in published studies (Table 1). The data selection process is unclear or not explained, which may explain the discrepancies in accuracy. Furthermore, it relies solely on laxity to establish a diagnosis with varying translation forces between 134N and 250N (Jenny et al., 2013; Klouche et al., 2015; Mouarbes et al., 2018). Measurements above 134 N proved impossible for most patients, making it difficult to interpret these measurements in a generalized way. Laxity is a parameter that can vary greatly depending on the configuration of the GNRB® (Bouguennec et al., 2015). Therefore, it is only possible to demonstrate the usefulness of the proposed solution based on our data.

However, it should be noted that all results were verified on 1384 different measurements with cross-validation and confirmed with 88 new patients. No other study has conducted such in-depth analyses and comparisons due to the limited availability of GNRB® measurements. For example, a comparative study between 49 (Di Iorio et al., 2014) and 118 (Klouche et al., 2015) patients will be inconclusive due to a large data gap.

The new diagnostic method based on this device and related analyses thus demonstrates the robustness and relevance of the solution for detecting and characterizing an ACL tear. Only the decision to retain values with tightening differences of more than 40 N, which is twice the limit value for accepting two laxity measurements as valid (Théo Cojean et al., 2023), may be open to discussion. The intention was to avoid having dominant parameters in the decision-making process, but their exclusion could potentially yield better results. Hence, it would be beneficial for all studies to share their data within a regulated research framework to establish a standardized test base. New DB could lead to a more coherent comparison of solutions and enable the development of a more effective diagnostic method with a diverse and comprehensive data set.

The association of two machine learning models with clear advantages is debatable, especially as the three thresholds have not been re-evaluated. A single model could have been sufficient to demonstrate the contribution of AI in diagnostics. However, this approach allows the creation of a new model that appears to be the best solution for leveraging dynamic laximetry device measurements such as GNRB, with gains of around 10%. Nevertheless, a significant improvement is possible by dynamically determining the three laxity thresholds to modify those used to reinforce diagnostic decisions.

It may be necessary to review the choice of hyperparameters, which are automatically selected to maximize the precision-to-loss ratio during training and testing. Although this choice allows for the quick addition of new data and evaluation of new sets, a manual or more statistical approach may increase the quality of diagnostics. Exploring the impact of all the hyperparameters suggested by Scikit-Learn and not just the most important ones (e.g. neighbor number for a Knn) is an area for further improvement.

The training dataset contains less than 30% of patients with rupture among the 1384 recorded data, which may explain why the model has an accuracy of approximately 62% for this type of injury. Therefore, obtaining more laxity measurements of patients with this type of rupture qualified by arthroscopy would be necessary. Increasing the data should result in a more accurate and robust model for a more comprehensive evaluation. The measurements from the other trials are not published, which is a limitation because collaboration could help improve this solution and find new ones. In the future, we need to find a way for each study to share its data following the rules of the ethics committees.

An alternative is to apply oversampling to increase the sample size could improve accuracy. The fear of creating bias in the database and lack of technical knowledge currently excludes artificial balancing of data choices. Another possible improvement would be to dynamically determine three laxity thresholds and modify those used as diagnostic aids.

Finally, all studies lack a “human” dimension that could be further explored, such as the physician’s opinion on knee instability. Adding other methods evaluated by physicians using Jerk, Lachman, or MRI results could also improve accuracy. There are various diagnostic methods, but they always are compared and never combined.

## 5. Conclusion

In conclusion, this study demonstrates that including only threshold parameter based on difference laxity measurements performed

at 134 N could be not enough and impact the diagnostic accuracy. The proposed solution counteract this bias and significantly improve the differentiation between healthy and torn ACLs by using a robust and accurate machine learning model in decision-making, including more parameters such as clamping force, sex, laxity results at different forces. Through a comprehensive analysis of a large dataset spanning multiple years, a more robust and accurate method (+6%) has been developed compared to static laxity interpretation. The choice of not directly confronting existing solutions or models, but proposing a bold combination, leads to the most robust and accurate diagnosis solution.

Due to its reduced cost and high precision, this dynamic laximetry device should be recommended as an alternative to MRI when used with our new diagnostic process. The results show that the solution is equivalent to this diagnostic standard and feasible everywhere due to the absence of radiation. However, operating on a patient solely based on a positive GNRB® or MRI test without considering their symptoms is not recommended, as it is possible to live with ACL tears. Therefore, we recommend using machine learning for diagnosis, combining the expertise and empathy of a physician that no algorithm can currently replicate.

## Code availability

The source code is strictly private and licensed.

## Funding

No funding was received for this study.

*Ethics approval*

Compliance with ethnical standard.

## CRediT authorship contribution statement

**Jean Mouchotte:** Development, Design, Writing of the article. **Matthieu LeBerre:** Editorial support, Knowledge exchange. **Théo Cojean:** Editorial support. **Henri Robert:** Database with ethic comities.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

I thank all the reviewers of the article as well as ESIEA and Genourob for their support.

## References

Alqahtani, Y., Murgier, J., Beaufils, P., Boisrenoult, P., Steltzlen, C., & Pujol, N. (2018). Anterior tibial laxity using the GNRB® device in healthy knees. *Knee*, *25*(1), 34–39.

Balasch, H., Schiller, M., Friebel, H., & Hoffmann, F. (1999). Evaluation of anterior knee joint instability with the rolimeter a test in comparison with manual assessment and measuring with the KT-1000 arthrometer. *Knee Surgery, Sports Traumatology, Arthroscopy*, *7*(4), 204–208.

Beaurain, F. (2020). GNRB (Medical device) vs MRI on anterior cruciate ligament (ACL) tears with arthroscopic validation. *International Journal of Research Studies Medical and Health Sciences*, *21*.

Bercovy, M., & Weber, E. (1995). Evaluation of laxity, rigidity and compliance of the normal and pathological knee. Application to survival curves of ligamentoplasties. *Revue de chirurgie orthopédique et réparatrice de l'appareil moteur*, *81*(2), 114–127.

Bohu, Y., Steltzlen, C., Lefevre, N., & Herman, S. (2012). Évaluation clinique d'une série continue de 55 cas de ligamentoplastie partielle du ligament croisé antérieur par la technique TLS®(greffe courte aux ischiojambiers). *Chir Orthop et Traumatol*, *98*(7), S373.

Bouguennec, N., Odri, G., Graveleau, N., & Colombet, P. (2015). Comparative reproducibility of TELOS™ and GNRB® for instrumental measurement of anterior tibial translation in normal knees. *Orthopeadics & Traumatology: Surgery & Reserach*, *101*(3), 301–305.

Branch, T. P., Mayr, H. O., Browne, J. E., Campbell, J. C., Stoehr, A., & Jacobs, C. A. (2010). Instrumented examination of anterior cruciate ligament injuries: minimizing flaws of the manual clinical examination. *Arthroscopy*, *26*(7), 997–1004.

Chang, M. J., Chang, C. B., Choi, J.-Y., Won, H. H., & Kim, T. K. (2013). How useful is MRI in diagnosing isolated bundle ACL injuries? *Clinical Orthopaedics and Related Research®*, *471*, 3283–3290.

Collette, M., Courville, J., Forton, M., & Gagniere, B. (2012). Objective evaluation of anterior knee laxity; comparison of the KT-1000 and GNRB® arthrometers. *Knee Surgery, Sports Traumatology, Arthroscopy*, *20*(11), 2233–2238.

Crawford, R., Walley, G., Bridgman, S., & Maffulli, N. (2007). Magnetic resonance imaging versus arthroscopy in the diagnosis of knee pathology, concentrating on meniscal lesions and ACL tears: a systematic review. *British Medical Bulletin*, *84*(1), 5–23.

Di Iorio, A., Carnesecchi, O., Philippot, R., & Farizon, F. (2014). Analyse multimodale des ruptures du ligament croisé antérieur: une étude prospective sur 49 cas. *R Chir Orthop et Traumatol*, *100*(7), 537–540.

Ebrahimipour, H., Mirfeizi, S. Z., Najar, A. V., Kachooei, A. R., Ariamanesh, A. S., Ganji, R., Esmaeeli, H., Salari, H., & Vejdani, M. (2014). Developing an appropriateness criteria for knee MRI using the rand appropriateness method (RAM)-2013. *Archives of Bone and Joint Surgery*, *2*(1), 47.

Gustafsson, T., Östenberg, A. H., & Alricsson, M. (2020). ACL diagnosis–The correlation between Rolimeter and MRI. *Sports Orthopaedics and Traumatology*, *36*(3), 278–283.

Hairy, P. (2021). Les réseaux de neurones et la data augmentation. *MetalBlog*.

Jenny, J. Y., Arndt, J., & Surgery-France, C. A. O. (2013). Anterior knee laxity measurement using stress radiographs and the GNRB® system versus intraoperative navigation. *Orthopeadics & Traumatology: Surgery & Reserach*, *99*(6), S297–S300.

Jenny, J. Y., Puliero, B., Schockmel, G., Harnoist, S., & Clavert, P. (2017). Experimental validation of the GNRB® for measuring anterior tibial translation. *Orthopaedics & Traumatology: Surgery & Research*, *103*(3), 363–366.

Klasan, A., Putnis, S. E., Kandhari, V., Oshima, T., Fritsch, B. A., & Parker, D. A. (2020). Healthy knee KT1000 measurements of anterior tibial translation have significant variation. *Knee Surgery, Sports Traumatology, Arthroscopy*, *28*(7), 2177–2183.

Klouche, S., Lefevre, N., Cascua, S., Herman, S., Gerometta, A., & Bohu, Y. (2015). Diagnostic value of the GNRB® in relation to pressure load for complete ACL tears: A prospective case-control study of 118 subjects. *Orthopeadics & Traumatology: Surgery & Reserach*, *101*(3), 297–300.

Learnia, M. (2022). Ensemble learning : Bagging, boosting et stacking (25/30). URL https://www.youtube.com/watch?v=7C_YpudYtw8.

Lefevre, N., Bohu, Y., Naouri, J., Klouche, S., & Herman, S. (2014). Validity of GNRB® arthrometer compared to Telos™ in the assessment of partial anterior cruciate ligament tears. *Knee Surgery, Sports Traumatology, Arthroscopy*, *22*(2), 285–290.

Mahesh, B. (2020). Machine learning algorithms-a review. *IJSR*, *9*, 381–386.

Micheo, W., Hernández, L., & Seda, C. (2010). Evaluation, management, rehabilitation, and prevention of anterior cruciate ligament injury: current concepts. *PM&R*, *2*(10), 935–944.

Mouarbes, D., Cavaignac, E., Chiron, P., Bérard, E., & Murgier, J. (2018). Evaluation of reproducibility of robotic knee testing device (GNRB) on 60 healthy knees. *Journal of Orthopaedics*, *15*(1), 94–98.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine*, *375*(13), 1216.

Pässler, H., & März, S. (1986). Der radiologische Lachman-Test—eine einfache und sichere Methode zum Nachweis von Kreuzbandschäden. *European Journal of Trauma*, *12*(6), 295–300.

Phelan, N., Rowland, P., Galvin, R., & O'Byrne, J. M. (2016). A systematic review and meta-analysis of the diagnostic accuracy of MRI for suspected ACL and meniscal tears of the knee. *Knee Surgery, Sports Traumatology, Arthroscopy*, *24*(5), 1525–1539.

Robert, H., Nouveau, S., Gageot, S., & Gagniere, B. (2009). A new knee arthrometer, the GNRB®: experience in ACL complete and partial tears. *Orthopeadics & Traumatology: Surgery & Reserach*, *95*(3), 171–176.

Saravia, A., Cabrera, S., Molina, C. R., Pacheco, L., & Muñoz, G. (2020). Validity of the Genourob arthrometer in the evaluation of total thickness tears of anterior cruciate ligament. *Journal of Orthopaedics*, *22*, 203–206.

Schuster, A. J., McNicholas, M. J., Wachtl, S. W., McGurty, D. W., & Jakob, R. P. (2004). A new mechanical testing device for measuring anteroposterior knee laxity. *The American Journal of Sports Medicine*, *32*(7), 1731–1735.

Théo Cojean, Cécile Batailler, Henri Robert, & Laurence Cheze (2023). GNRB® laximeter with magnetic resonance imaging in clinical practice for complete and partial anterior cruciate ligament tears detection: A prospective diagnostic study with arthroscopic validation on 214 patients.

Torg, J. S., Conrad, W., & Kalen, V. (1976). Clinical I diagnosis of anterior cruciate ligament instability in the athlete. *The American Journal of Sports Medicine*, *4*(2), 84–93.

Van Dyck, P., Vanhoenacker, F. M., Gielen, J. L., Dossche, L., Van Gestel, J., Wouters, K., & Parizel, P. M. (2011). Three tesla magnetic resonance imaging of the anterior cruciate ligament of the knee: can we differentiate complete from partial tears? *Skeletal Radiology*, *40*, 701–707.

Zhao, M., Zhou, Y., Chang, J., Hu, J., Liu, H., Wang, S., Si, D., Yuan, Y., & Li, H. (2020). The accuracy of MRI in the diagnosis of anterior cruciate ligament injury. *Annals of Translational Medicine*, *8*(24).

Zhou, Z. H. (2021). *Machine learning*. Springer Nature.