Predicting ACL Injuries Using Machine Learning Models and Tibial Anatomical Predictors

by

Cheng-Hao (Howard) Kao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2025

© Cheng-Hao (Howard) Kao 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The tibial slope and the tibial depth are well-established risk factors for Anterior Cruciate Ligament (ACL) injury. As ML continues to progress, it has become an increasingly reliable tool for clinical screening and risk factor analysis. This thesis aims to develop and validate an explainable prognostic ML model to predict ACL injury outcomes from these Tibial Anatomical Feature (TAF), and identify the most predictive features among these parameters.

A dataset comprising Coronal Tibial Slope (CTS), Medial Tibial Slope (MTS), Lateral Tibial Slope (LTS), Medial Tibial Depth (MTD), and sex was constructed using MRI scans taken from 104 subjects (44 males: 22 injured, 22 uninjured; 60 females: 27 injured, 33 uninjured). Two distinct ML pipelines were developed: a self-developed pipeline (including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), XGBoost, CATBoost, Multi-Layer Perceptron (MLP), and TabNet) and an advanced AutoGluon pipeline (including XGBoost, LightGBM, CatBoost, TabPFN, TabM, TabICL, MITRA, and their weighted ensembles). Both were designed as end-to-end pipelines to process the dataset and output predictions with integrated feature importance explanations. Empirically, the AutoGluon Pipeline demonstrated superior performance and training-time efficiency. The recommended *F2-tuned standard ensemble* achieved an F2-score of 0.736 on the validation set. On the test set, it demonstrated a test balanced accuracy of 0.955, F1-score of 0.952, F2-score of 0.980, ROC AUC of 1.000, precision of 0.909, and recall of 1.000. A full-dataset model, the *F2-tuned full-dataset ensemble* refitted on the entire dataset for clinical deployment achieved a validation F2-score of 0.813. The global feature importance analyses performed via SHapley Additive exPlanations (SHAP), established the descending order of influences as MTD, LTS, MTS, CTS, and sex.

In summary, the study recommends two versions of the F2-tuned prognostic models, one being a standard ensemble model and the other a full-dataset ensemble. The former, which demonstrated moderately high predictive power, was designed for subsequent research comparison. The latter, without access to the original held-out test set, is constructed for maximum robustness and generalization in real-life clinical deployment. Global feature importance analyses elucidated from the standard ensemble decreased MTD along with increased LTS and MTS as most contributive features for ACL injury. These models serve as both feature attribution tools as well as clinical screening tools. These models are intended to be integrated into clinical practice as explainable machines to assist clinicians in predicting the likelihood of ACL injury.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Professor Naveen Chandrashekar, for his maximum support, mentorship, and acknowledgment throughout my master's degree. His guidance was invaluable.

I am grateful to Professor Javad Hashemi for generously providing the dataset, and to Dr. James R. Slauterbeck and all the participants from the University of Vermont–associated hospital whose contributions made this dataset possible.

I extend my sincere thanks to Daniel Loewen, Professor Cederick Landry, and Professor Brendan L. Pinto for their valuable collaboration, feedback, and support in the original direction of this thesis and other related projects.

I would also like to thank my committee members, Professor William Melek and Professor HJ Kwon, for their insightful feedback and advice throughout the review process.

Finally, I thank my family, lab members, and close friends for their unconditional support before and during my pursuit of a graduate degree.

Dedication

This is dedicated to the inspirational minds, my family, and dear friends who have been part of this journey. Thank you for your presence in my life; I genuinely appreciate each and every single one of you.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

Dedication v

List of Figures x

List of Tables xiii

List of Abbreviations xv

1 Overview 1

1.1 Motivation 1

1.2 Objective 1

1.3 Thesis Outline 2

2 Background and Related Work 3

2.1 ACL Injury Risk Factor 3

2.1.1 Categories 3

2.1.2 Measurement Techniques 6

2.1.3 Pre-processing . . . 14
2.2 Machine Learning in ACL Injury . . . 14
2.3 ML Models for Small Tabular Datasets . . . 16
2.3.1 Instance-Based Methods . . . 17
2.3.1.1 KNN . . . 17
2.3.2 Kernel-Based Methods . . . 18
2.3.2.1 SVM . . . 18
2.3.3 GBT . . . 19
2.3.3.1 RF . . . 19
2.3.3.2 XGBoost . . . 20
2.3.3.3 CatBoost . . . 20
2.3.3.4 LightGBM . . . 21
2.3.4 DNN . . . 21
2.3.4.1 MLP . . . 21
2.3.5 TabNet . . . 22
2.3.5.1 TabM . . . 23
2.3.5.2 TabPFN v2 . . . 24
2.3.5.3 TabICL . . . 25
2.3.5.4 MITRA . . . 27
2.4 Gaps in Current Research . . . 27

3 Methodology **29**
3.1 Dataset . . . 29
3.1.1 Source . . . 29
3.1.2 Feature Selection . . . 32
3.1.3 Data Synthesis . . . 32
3.1.4 Metric Selection . . . 32
3.2 Self-Developed Pipeline . . . 33

3.2.1 Training and Tuning . . . 33
3.2.2 Evaluation . . . 38
3.3 AutoGluon Pipeline . . . 38
3.3.1 Training and Tuning . . . 38
3.3.2 Evaluation . . . 41
3.4 Additional Analyses . . . 41
3.4.1 Prediction Confidence and Feature Importance Explanation . . . 42
3.4.2 Full-dataset Model . . . 43

4 Results **44**
4.1 Self-Developed Pipeline . . . 44
4.1.1 Test Performance . . . 44
4.1.1.1 Accuracy-Tuned TabNet . . . 44
4.1.1.2 F1-Tuned TabNet . . . 46
4.1.2 Global Feature Importance . . . 48
4.1.2.1 Accuracy-Tuned TabNet . . . 49
4.1.2.2 F1-Tuned TabNet . . . 50
4.2 AutoGluon Pipeline . . . 52
4.2.1 Test performance . . . 52
4.2.1.1 Balanced-Accuracy-Tuned Standard Ensemble . . . 52
4.2.1.2 F1-Tuned Standard Ensemble . . . 55
4.2.1.3 F2-Tuned Standard Ensemble . . . 58
4.2.2 Global Feature Importance . . . 61
4.2.2.1 Balanced-Accuracy-Tuned Standard Ensemble . . . 62
4.2.2.2 F1-Tuned Standard Ensemble . . . 64
4.2.2.3 F2-Tuned Standard Ensemble . . . 66
4.3 Full-dataset Model . . . 68

5 Discussion **72**
5.1 Results Summary . . . 72
5.2 Model Choice . . . 73
5.3 Pipeline Comparison . . . 74
5.3.1 Ground Truth Distribution on Lower-Dimensional Simplex . . . 74
5.3.1.1 Self-Developed Pipeline . . . 75
5.3.1.2 AutoGluon Pipeline . . . 78
5.3.2 Global Feature Explanation Stability . . . 82
5.3.2.1 Self-Developed Pipeline . . . 82
5.3.2.2 AutoGluon Pipeline . . . 85
5.4 Literature Comparison . . . 87
5.4.1 Predictive Performance . . . 87
5.4.2 Feature Importance and Potential Risk Factors . . . 87
5.5 Local Feature Importance . . . 89
5.6 Full-Dataset Ensemble . . . 93
5.7 Model Suitability . . . 94
5.7.1 Self-Developed Pipeline . . . 94
5.7.2 AutoGluon Pipeline . . . 95

6 Conclusion **98**

7 Limitation and Future Work **99**

References **101**

APPENDICES **116**

A Raw Dataset **117**

List of Figures

2.1 Anatomic structure of the ACL [123] 4
2.2 Coronal plane represented by line AA [41]. 7
2.3 Coronal tibial longitudinal axis [41] 7
2.4 Coronal tibial slope with respect to the normal reference axis [41] 8
2.5 Medial plane represented by line AA, lateral plane represented by line BB, and sagittal planes for determining longitudinal axes represented by dotted lines [41]. 9
2.6 Medial tibial longitudinal axis [41] 9
2.7 Medial tibial slope with respect to the normal reference axis [41] 10
2.8 Lateral tibial slope with respect to the normal reference axis [41] 11
2.9 Medial tibial depth represented by line distance between lines AA and BB [41] 11
2.10 Central longitudinal axis [54] 12
2.11 Medial tibial slope with respect to reference axis [54] 13
2.12 Lateral tibial slope with respect to reference axis [54] 13
2.13 Raw dataset correlation heatmap 14
2.14 An example of 3-nearest neighbor, where q_1 would be classified as "o" and q_2 as "x" [22]. 18
2.15 An example of linear SVM [72]. Depending on the kernel chosen, the decision boundary can take different forms 19
2.16 Decision tree architecture [72]. A random forest is the ensemble of an arbitrary number of such decision trees, serving as a voting committee 20

2.17 MLP architecture [72]. Initially designed to mimic human nervous systems 21
2.18 TabNet Architecture, designed by Google [6]. The Upper stream yields predictions while the bottom stream provides feature importance 22
2.19 TabNet feature and attentive tranformers [6]. Despite the name *transformer*, they are unrelated to the famous self-attention-based transfomer 23
2.20 TabNet encoder-decoder structure for self-supervised learning [6] 24
2.21 TabM architecture [37] 24
2.22 TabPFN architecture [121]. This architecture is different from TabPFN v1 25
2.23 TabICL TF_{col} architecture [91] 26
2.24 TabICL TF_{col} sub-architecture [91] 26
2.25 TabICL TF_{icl} architecture [91] 27

3.1 2D UMAP projection of raw data 31
3.2 3D UMAP projection of raw data 31
3.3 Learning paradigm comparison between two pipelines 33
3.4 Preliminary results manifesting overfitting 35
3.5 An example of 5-fold cross-validation [87] 36
3.6 Bayesian optimization history 37
3.7 K-fold bagging [26] 39
3.8 Multi-layer stacking [26] 40

4.1 Confusion for accuracy-tuned TabNet 46
4.2 Confusion for F1-tuned TabNet 48
4.3 Beeswarm plot for accuracy-tuned TabNet 49
4.4 Internal feature importance of accuracy-tuned TabNet 50
4.5 Beeswarm plot for F1-tuned TabNet 51
4.6 Internal feature importance of F1-tuned TabNet 51
4.7 Confusion matrix for the balanced-accuracy-tuned ensemble 55
4.8 Confusion matrix for the F1-tuned ensemble 58

4.9 Confusion matrix for the F2-tuned ensemble 61
4.10 Averaged global feature importance by balanced-accuracy-tuned ensemble . 62
4.11 Combined view of global feature importance by balanced-accuracy-tuned ensemble 63
4.12 Averaged global feature importance by F1-tuned ensemble 64
4.13 Combined view of global feature importance by F1-tuned ensemble 65
4.14 Averaged global feature importance by F2-tuned ensemble 66
4.15 Combined view of global feature importance by F2-tuned ensemble 67
4.16 Combined view of global feature importance for balanced-accuracy-tuned ensemble 69
4.17 Combined view of global feature importance for F1-tuned ensemble 70
4.18 Combined view of global feature importance for F2-tuned ensemble 71

5.1 Ground truth distribution on 1D simplex for accuracy-tuned TabNet . . . 76
5.2 Ground truth distribution on 1D simplex for F1-tuned TabNet 77
5.3 Ground truth distribution on 1D simplex for balanced-accuracy-tuned ensemble 79
5.4 Ground truth distribution on 1D simplex for F1-tuned ensemble 80
5.5 Ground truth distribution on 1D simplex for F1-tuned ensemble 81
5.6 Feature importance comparison for accuracy-tuned TabNet 82
5.7 Feature importance comparison for F1-tuned TabNet 83
5.8 Comparison of global SHAP values across two TabNet models 84
5.9 Mean global SHAP values across tuning metrics for more stable ranking . . 84
5.10 Comparison of global SHAP values across three ensemble models 86
5.11 Mean global SHAP values across tuning metrics for more stable ranking . . 86
5.12 Combined view of global feature importance for balanced-accuracy-tuned ensemble 90
5.13 Combined view of global feature importance for F1-tuned ensemble 91
5.14 Combined view of global feature importance for F2-tuned ensemble 92

List of Tables

2.1 Bi-axial risk factor categorization matrix 5

3.1 Mean and standard deviation statistics for the injured and uninjured groups of male and female participants, computed from the entire dataset. They are expressed in the format of *mean* ± *standard deviation* 29

3.2 Change in validation and test performance when homogeneous bagging is applied 37

4.1 Best-performing models with the highest validation accuracies from each model family and their corresponding feature subsets. The integers 1 to 5 in the table represent different features - 1:CTS, 2: MTS, 3: LTS, 4:MTD, 5:sex 45

4.2 Accuracy-tuned TabNet hyperparameters 45

4.3 Best-performing models with the highest validation F1-scores from each model family and their corresponding feature subsets. The integers 1 to 5 in the table represent different features - 1:CTS, 2: MTS, 3: LTS, 4:MTD, 5:sex 47

4.4 F1-tuned TabNet hyperparameters 47

4.5 Performance of balanced-accuracy-tuned models ranked by validation score. Metrics are averaged across ten runs 53

4.6 Performance metrics for F2-tuned Ensemble different random seeds 54

4.7 Performance of F1-tuned models ranked by validation score. Metrics are averaged across ten runs 56

4.8 Performance metrics for F1-tuned Ensemble different random seeds 57

4.9 Performance of F2-tuned ranked by validation score. Metrics are averaged across ten runs . . . 59
4.10 Performance metrics for different random seeds (F2-focused) . . . 60
4.11 Validation performance of full-dataset models . . . 68

5.1 Best and average F1-tuned model ranking . . . 94
5.2 Best and average F1-tuned model ranking . . . 94
5.3 Best and average balanced-accuracy-tuned model ranking . . . 95
5.4 Best and average F1-tuned model ranking . . . 95
5.5 Best and average F2-tuned model ranking . . . 95
5.6 Best and average balanced-accuracy-tuned full-dataset model ranking . . . 96
5.7 Best and average F1-tuned full-dataset model ranking . . . 96
5.8 Best and average F2-tuned full-dataset model ranking . . . 97

A.1 TAFs for ACL injury prediction [41] . . . 117
A.1 TAFs for ACL injury prediction [41] . . . 118
A.1 TAFs for ACL injury prediction [41] . . . 119
A.1 TAFs for ACL injury prediction [41] . . . 120

List of Abbreviations

ACL Anterior Cruciate Ligament iii

ACLR Anterior Cruciate Ligament Reconstruction 1

CTS Coronal Tibial Slope iii

DNN Deep Neural Network 16

FC Fully Connected Layers 22

FIA Feature Importance Analysis 15

GBT Gradient Boosted Tree 16

ICL In-Context Learning 24

KNN K-Nearest Neighbor iii, 17

L1 Layer 1 52

L2 Layer 2 95

LTS Lateral Tibial Slope iii

MLP Multi-Layer Perceptron iii

MTD Medial Tibial Depth iii

MTS Medial Tibial Slope iii

OOF Out-of-Fold 40

PCA Principal Component Analysis 30

PFN Prior-Data Fitted Network 24

RF Random Forest iii

RFE Recursive Feature Elimination 32

RoPE Rotary Positional Encoding 25

SHAP SHapley Additive exPlanations iii

SOTA State-of-the-Art 16

SVM Support Vector Machine iii, 18

TAF Tibial Anatomical Feature iii

Chapter 1

Overview

1.1 Motivation

More than 120,000 ACL injuries are estimated to occur annually in the United States [78]. These injuries account for approximately half of all knee-related injuries in orthopaedic sports disorders [79, 83]. Traditionally, these injuries have been diagnosed through arthroscopy and MRI [65, 68]. Following diagnosis, Anterior Cruciate Ligament Reconstruction (ACLR) is often recommended as a prevalent surgical intervention in clinical settings, which has been shown to improve knee functionality and quality of life compared to non-surgical treatments especially in young athletic populations [29, 38]. However, patients remain susceptible to persistent post-injury complications from ACL injury such as osteoarthritis, asymmetrical strength, and residual pain (re-injury to ipsilateral or injury to contralateral knee), underscoring the need for enhanced prognostic and preventive strategies [29, 38]. Therefore, prevention of ACL injury is important.

1.2 Objective

The main objectives of this thesis are to (I) develop and validate classification ML models capable of reliably predicting ACL injury using TAFs, (II) elucidate the most informative features associated with ACL injuries, (III) create full-dataset variants for maximum capacity and robustness, and (IV) enable interpretable predictions for clinical use after deployment. In simplified terms, the aim of the thesis is to construct high-performance ML models using a clinical dataset and leverage explainable AI methods for identifying

top predictors among features selected, such that the models can contribute to the body of knowledge regarding ACL while assisting and even replacing medical professionals.

1.3 Thesis Outline

The thesis is structured as follows: Chapter 2 will provide a description of essential concepts, current status in literature related to risk factors for ACL injuries, and latest progress in ML architectures developed for small tabular dataset problems. Chapter 3 will cover the dataset properties and pre-processing, the two pipelines involved in the experimentation and the intricacies in them, and the follow-up analyses conducted on built models. Chapter 4 will present major findings from the experiments. Chapter 5 will examine the results in depth and compare them with existing literature. Chapter 6 will summarize the key findings from this thesis. Chapter 7 will highlight the existing limitations in this work and identify future directions.

Chapter 2

Background and Related Work

2.1 ACL Injury Risk Factor

2.1.1 Categories

The ACL is a key stabilizing ligament of the knee joint, connecting the anterior intercondylar area of the tibia and the posteriomedial aspect of the lateral femoral condyle as shown in Figure 2.1. It is a critical component of the knee that restricts internal rotation and anterior tibial translation [123].

Injury to the ACL is the most prevalent type of knee-related injuries, especially among athletic populations engaged in sports requiring frequent jumping, pivoting, and rapid decelerating [35, 74]. These injuries are classified among multiple axes. Mechanistically, they are divided into contact and non-contact ACL injuries. Contact injuries result from direct external impact to the knee, whereas non-contact injuries, which represent the majority of cases, arise from a wide variety of other factors [88]. Chronologically, ACL injuries can be categorized as primary or secondary. Primary injuries represent the first occurrence of ACL injuries, while secondary ACL injuries represent the re-injury after reconstruction treatment. Both are extensively studied in literature [13, 57, 46]. Further categorization axes, such as severity (partial vs. complete) [28] and complexity (isolated vs. combined) [94] are less emphasized in literature.

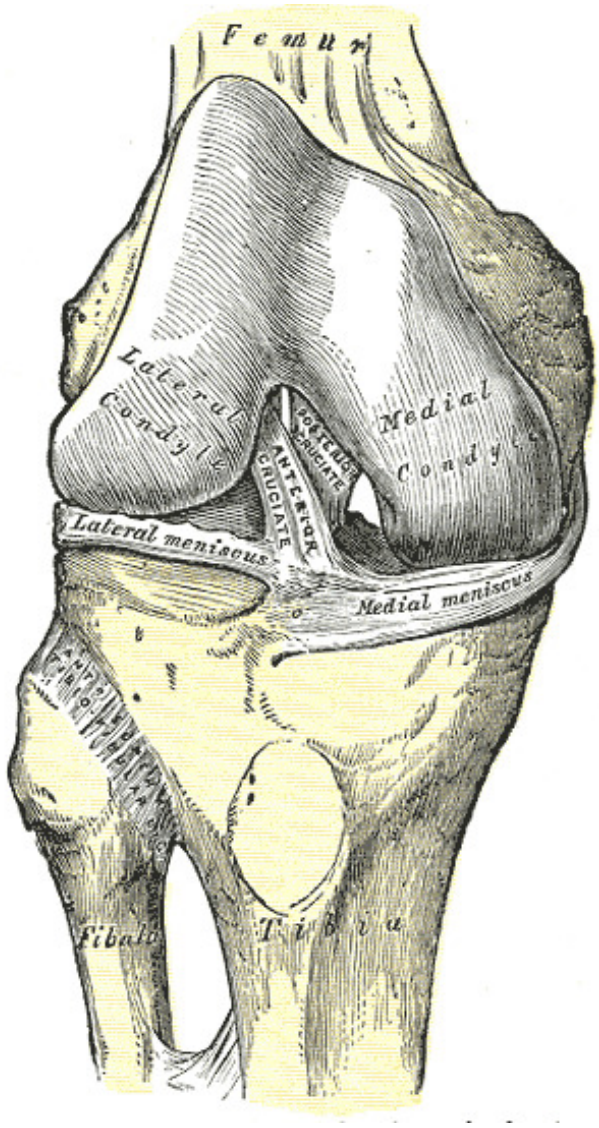


Figure 2.1: Anatomic structure of the ACL [123]

The wide range of risk factors implicated in ACL injuries are systematically organized along two different axes: the origin paradigm and the intervention paradigm. The origin paradigm categorizes risk factors as intrinsic or extrinsic, depending on whether the source of risk arises from within the patient or from external circumstances [102]. The intervention paradigm, on the other hand, categorizes risk factors as modifiable or non-modifiable, based on the degree to which an individual can influence or alter the factor [88]. The

intervention paradigm, although commonly used to subdivide intrinsic risk factors, can be applied to extrinsic factors as well. Together, these two paradigms form a bi-axial risk-factor categorization matrix, enabling clearer visualization:

Table 2.1: Bi-axial risk factor categorization matrix

Category	Non-Modifiable	Modifiable
Intrinsic	Anatomical Geometry Biological Sex Injury History	Muscular Strength Kinematic Failure Patterns Body Mass Index
Extrinsic	Weather Conditions Environmental Conditions	Workload Management Footwear Selection Technique Training

Different studies often choose to include varying sets of potential risk factors based on their research focus and data availability. However, the majority of risk factors studied tend to be intrinsic. This can be attributed to three different reasons: (I) reliable measurement, (II) direct causal relevance, and (III) ethical limitations. Reliable measurement refers to the relative ease with which consistent and reproducible measurements of intrinsic factors can be attained. Extrinsic factors are far more difficult to reliably measure due to their complexity and variability which are often dependent on dynamic external conditions. Direct causal relevance refers to the direct impact intrinsic risk factors have on developed conditions. Since ACL injuries are developed within the biological structure of the knee, intrinsic factors are often the primary cause and provide more interpretability for research verification. External factors, in contrast, more commonly act as triggers or preconditions that interact with preexisting intrinsic vulnerabilities that cause biomechanical failure. Ethical limitation describes the difficulties present in conducting experiments or studies with extrinsic factors, as it is unethical to request participants to partake in dangerous scenarios and expose them to known or potential risks. As a result, intrinsic risk factors for ACL such as parental ACL injury history, knee stability, sex-specific anatomical differences, genetic factors, hormones, bony morphology, etc., are more frequently investigated in both observational and experimental studies [11, 46, 106, 103, 12].

Tibial slopes and depths, a specific subset of bony morphology, have been gaining attention in ACL injury research as major risk factors for ACL injury and ACLR rupture [125]. These variables, in general, describe the sagittal inclination and convexity of different regions in the tibial plateau [61, 125]. This research focus was formally established by the Hashemi et al., which clearly defined and measured the distinct tibial slope components, including MTS, LTS, and CTS, using MRI scans [42]. Their work standardized one of the most universal measurement techniques for these important tibial morphologic variables.

The clinical investigation of the relationship between TAFs and ACL injuries remains active, as evidenced by recent literature [46, 10, 104].

2.1.2 Measurement Techniques

The most popular choices of tibial anatomical measurements are well-described in [125]. These techniques have been classified as the *Hashemi*, *Hudek*, *Longitudinal*, *Lateral mechanical axis*, and *other* (*Brandon*, *Hohmann* and *Hendrix*) methods. Among the techniques applied to ACL injury and ACL graft failure analyses using the MRI modality, the *Hashemi* and *Hudek* methods are the most commonly adopted. The main difference between the two is their approach of defining the tibial longitudinal axis, which serves as the critical reference for posterior tibial slope measurements. The *Hashemi* method determines this axis by identifying the midpoints of parallel lines drawn between anterior and posterior cortices at two different height (one proximal and one distal) and connecting these midpoints [42]. contrarily, the *Hudek* method defines the axis by locating the centers of circles drawn at two levels (one proximal *cranial* and one distal *caudal*) and the center [54].

The *Hashemi* method is detailed as follows:

1. Locate the transverse plane at the tibiofemoral joint
2. CTS measurement
   (a) Locate the coronal plane that passes through the tibial plateau centroid (Figure 2.2)
   (b) Construct a longitudinal axis by connecting middle points of two distal horizontal lines drawn between the medial and lateral cortices (Figure 2.3)
   (c) Construct a normal reference axis orthogonal to the longitudinal axis (vector p in Figure 2.4)
   (d) Construct a target line between the peaks of the medial and lateral plateau (line AB Figure in 2.4)
   (e) Measure the angle of the target line with respect to the normal axis

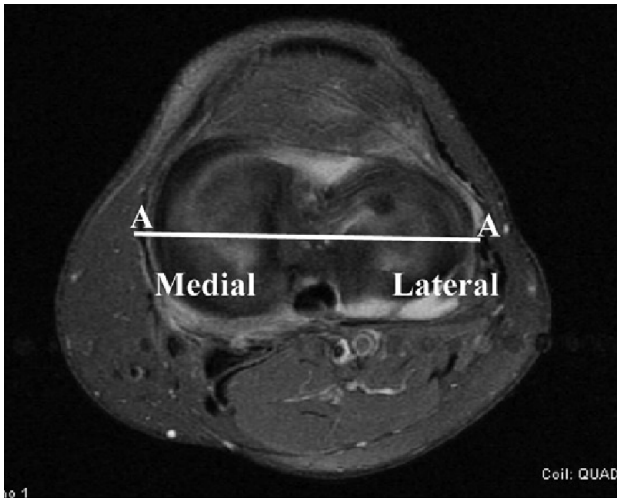


Figure 2.2: Coronal plane represented by line AA [41].

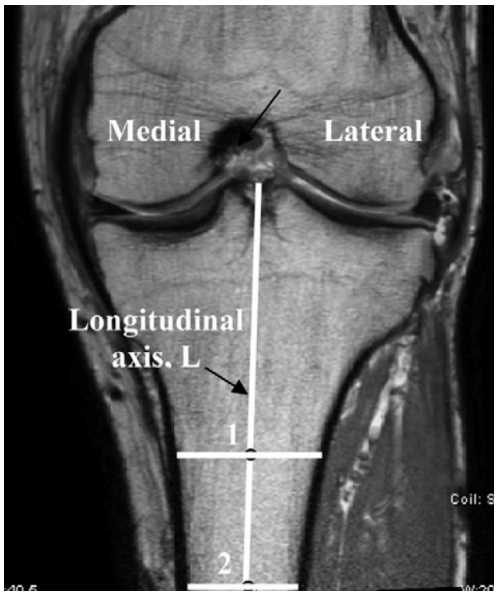


Figure 2.3: Coronal tibial longitudinal axis [41]



Figure 2.4: Coronal tibial slope with respect to the normal reference axis [41]

3. MTS measurement

   (a) Locate two sagittal planes, one for determining the longitudinal axis (left dotted line in Figure 2.5) and the other for measuring the MTS (line AA in Figure 2.5), respectively

   (b) Construct a longitudinal axis by connecting middle points of two distal horizontal lines drawn between the anterior and posterior cortices on the former sagittal plane (Figure 2.6)

   (c) Construct a normal reference axis orthogonal to the longitudinal axis (vector p in Figure 2.7)

   (d) Construct a target line connecting the peaks of the medial plateau on the latter sagittal plane (line AB Figure 2.7)

   (e) Measure the angle of the target line with respect to the normal axis

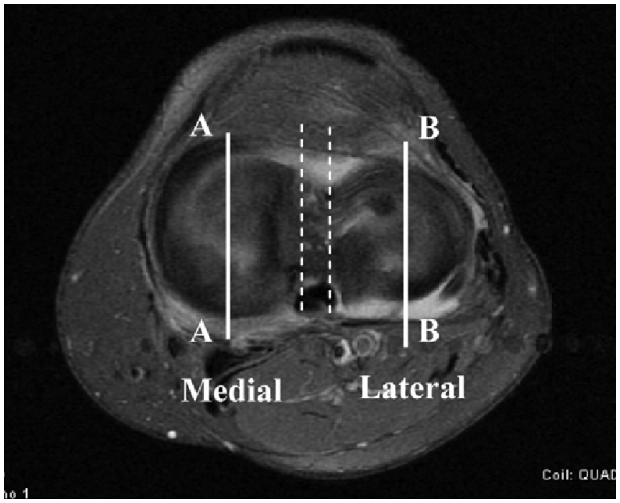


Figure 2.5: Medial plane represented by line AA, lateral plane represented by line BB, and sagittal planes for determining longitudinal axes represented by dotted lines [41].

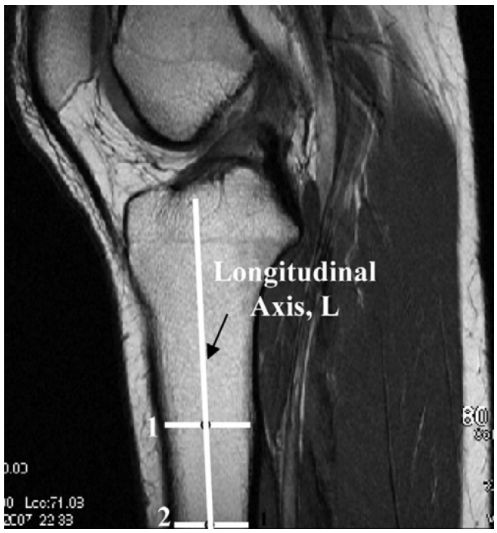


Figure 2.6: Medial tibial longitudinal axis [41]

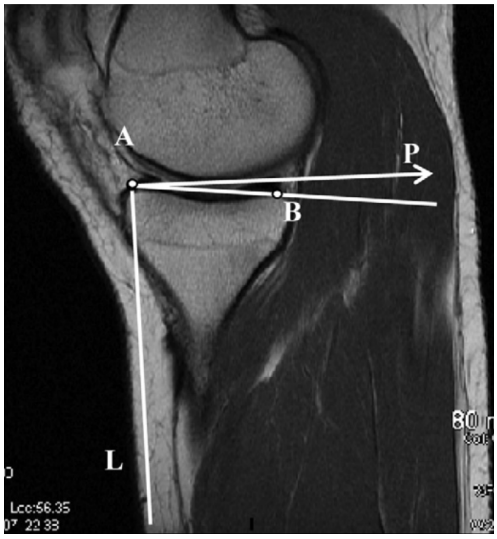


Figure 2.7: Medial tibial slope with respect to the normal reference axis [41]

4. LTS measurement

   (a) Locate two sagittal planes, one for determining the longitudinal axis (right dotted line in Figure 2.5) and the other for measuring the LTS (line BB in Figure 2.5), respectively

   (b) Construct a longitudinal axis by connecting middle points of two distal horizontal lines drawn between the anterior and posterior cortices on the former sagittal plane (Figure 2.6)

   (c) Construct a normal reference axis orthogonal to the longitudinal axis (vector p in Figure 2.8)

   (d) Construct a target line connecting between the peaks of the lateral plateau on the latter sagittal plane (line AB in Figure 2.8)

   (e) Measure the angle of the target line with respect to the normal axis

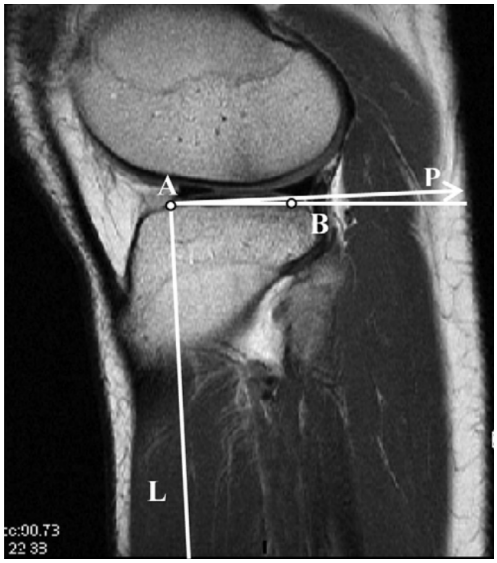


Figure 2.8: Lateral tibial slope with respect to the normal reference axis [41]

5. MTD measurement

   (a) Reuse the plane for MTS measurement obtained in step 3a (Figure 2.5)

   (b) Construct a target line connecting the peaks of the medial plateau on the plane (line AA in Figure 2.9)

   (c) Construct a reference axis parallel to the target line that passes through the lowest point of the medial plateau (line BB in Figure 2.9)

   (d) Measure the line distance between the target line and the reference axis

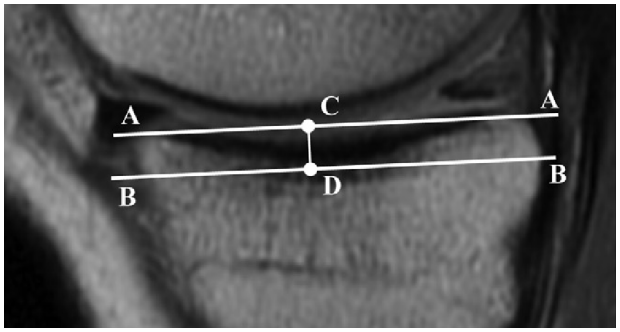


Figure 2.9: Medial tibial depth represented by line distance between lines AA and BB [41]

The *Hudek* method follows nearly identical procedures for measuring MTS and LTS, performing the same steps 3 and 4 for MTS and LTS measurements, with the only exceptions being steps 3b and 4b. Instead of defining the tibial axes with respect to the midpoint of parallel lines, it leverages circles drawn at two levels on the central slice of MRI slice, as depicted in Figure 2.10. Similar to the *Hashemi* method, tibial slopes can be measured against the normal reference axes defined by the tibial longitudinal axis (Figures 2.11 and 2.12).

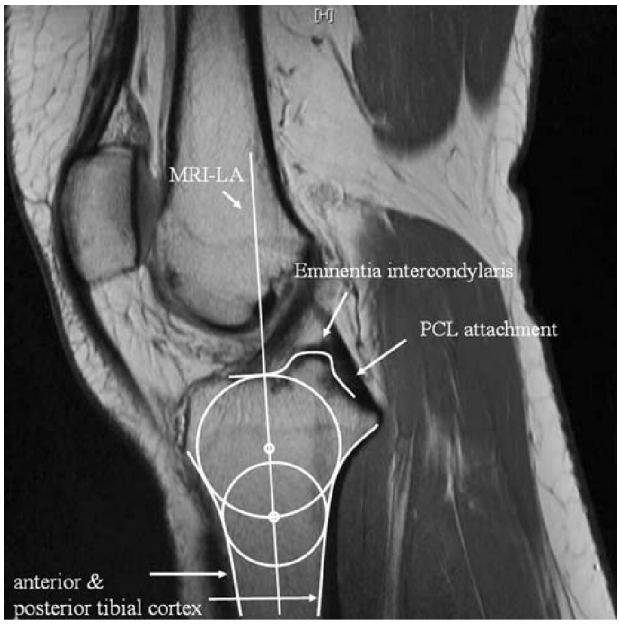


Figure 2.10: Central longitudinal axis [54]

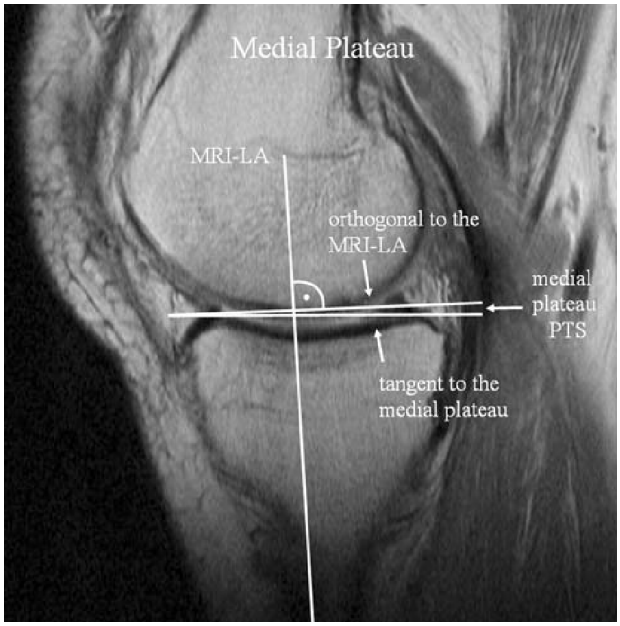


Figure 2.11: Medial tibial slope with respect to reference axis [54]

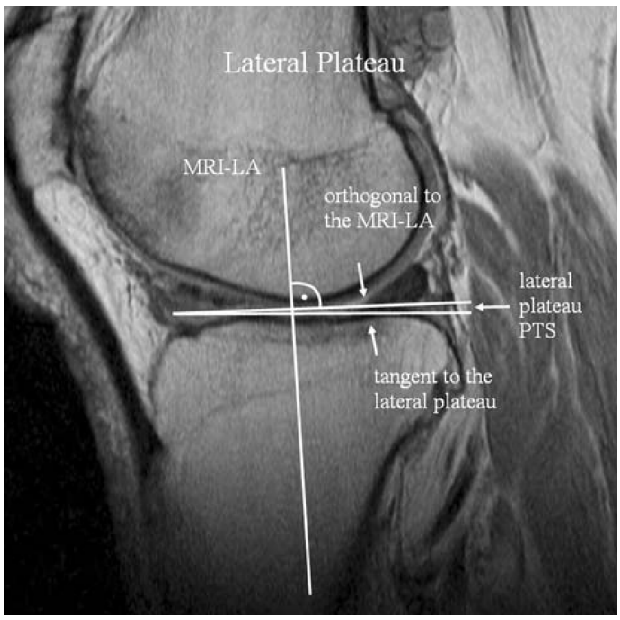


Figure 2.12: Lateral tibial slope with respect to reference axis [54]

2.1.3 Pre-processing

The dataset was obtained in a tabular, unimodal format from [41]. There were no missing entries, and the class distribution was approximately balanced. No strong correlations were observed among feature pairs or between features and the target variable, with the highest absolute correlation coefficient being 0.46, as shown in Figure 2.13.

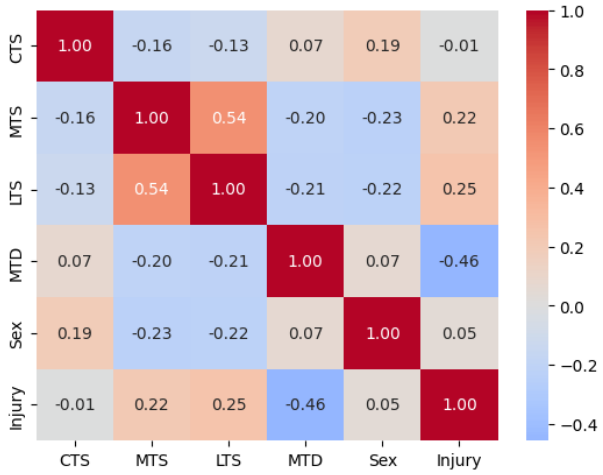


Figure 2.13: Raw dataset correlation heatmap

2.2 Machine Learning in ACL Injury

Recently, ML techniques have emerged as a promising complement, leveraging classical or deep learning models to achieve high diagnostic and prognostic accuracies by learning patterns from existing datasets [7, 97, 64, 60, 5]. They are increasingly recognized for their ability to handle complex, high-dimensional data and uncover intricate relationships that may not be readily discernible for conventional statistical methods [67, 24]. Many ML models leverage large datasets to learn structures and make predictions without assuming a priori functional forms, enhancing flexibility in risk assessment [92]. Additionally, carefully designed ML models are less prone to human errors, which makes them a reliable option for assisting or replacing a medical professional's judgment [20, 13]. Faltstrom et al. combined clinical assessments, functional performance exams, and psychological factors in their dataset to predict female soccer players at risk of ACL re-injury, which resulted in 100% accuracy, 89% sensitivity, and 76% specificity [33]. Schilaty et al., as another

example, focused on analysing the anterior tibial shear, knee abduction moment, vertical ground-reaction force, and internal tibial rotation using linear discriminant analysis. They successfully predicted injury outcomes in different components of the knee, which achieved 94.1% accuracy and 98.0% AUC [96].

By integrating feature importance analyses conducted on such predictive models with insights gained from traditional methods, risk factors are better identified, enabling researchers to work towards developing more effective and comprehensive strategies for ACL injury mitigation. Bailey et al. used a convolutional-neural-network-based approach for prediction of ACL re-injury. Through Feature Importance Analysis (FIA), they identified multiple bone shape characteristics contributing to ACL re-injury [8]. Lopez et al. deployed an artificial neural network and showed that it significantly outperformed linear regression in the tasks of predicting ACLR readmissions and reoperation. They further concluded surgery conditions, operative time, body mass index, age, and race as top indicators [71]. Lu et al. performed a comparative analysis among five different ML models to predict secondary ACL injury, resulting in accuracies of roughly 70%. They extracted the most contributive features, such as body mass index, age at injury, and history of arthroscopies on the contralateral knee, etc. [75]. Another study by the MARS Group deployed an AutoPrognosis model and reported femoral and tibial tunnel size and position in ACLR as well as the use of allograft in revision ACLR as primary risk factors [110]. Johnson et al. implemented eight separate ML models for predicting post-ACLR reoperations and found inflammatory conditions, distal tears, concurrent medial collateral ligament treatment, visual analog scale, and proximal tears to be the primary risk factors using the XGBoost model [55]. Taborri et al. experimented with nine different classifiers in order to predict the Landing Error Score of female basketball players, which is associated with risk of ACL injury. The best model from this study, a linear support vector machine, elucidated ellipse area, load absorption-related parameters, and leg mobility as the top predictors [107].

Tibial slopes and depths, having been identified in recent literature as predictive indicators and major risk factors for ACL injury and ACLR rupture [125], have been included in an increasing number of ML analysis frameworks for enhancing prognostic performance. For instance, Zhang et al., in a study involving a weighted-average voting ensemble model that predicts the risk of post-ACLR graft failure, found posterior slope to be one of the top five predictors [126]. Ye et al. concluded by comparing six distinct ML models that posterior tibial slopes, along with medial meniscal removal, participation in competitive sports, etc., were the most predictive indicators predicting post-operative graft failures [122]. The studies above, combined with increasing emphasis on the tibial slopes and depths [10, 88], have raised the potential of the tibial structure in clinical ACL injury prediction.

2.3 ML Models for Small Tabular Datasets

In order to maintain the interpretability of ML models in the field of medicine, researchers often choose to employ heterogeneous tabular datasets which offer inherent flexibility and can encapsulate various types of information of greater medical significance [31]. Such data format enables the semantic separation of features, providing an interpretable representation of complex variables. For instance, while a MRI scan of a knee tibial conveys more intuitive visual information, tabulated features like various tibial slopes can serve as a more interpretable and medically relevant representation. Additionally, tabulated data allows for the combination of diverse elements, such as self-reported scores alongside objective measurements, resulting in a more comprehensive dataset. Tabulated datasets, however, present several challenges for ML applications. One key challenge is that they are often noisy due to variability in medical equipment and measurement processes, to which not all models are robust [128]. This is especially problematic for models like Deep Neural Network (DNN)s as the majority of them are designed to handle homogeneous data rather than heterogeneous structures with dense numerical and sparse categorical features [15, 58]. Studies have pointed out the effectiveness of employing ensemble techniques, particularly bagging, to overcome such issues [100]. Another challenge is that structural patterns or invariances in tabular data are considerably more difficult to identify compared to those in other data structures. Without exploitable inductive biases, such as spatial relationships used by the convolutional neural network and vision transformer families [82], model architecture adjustment and data augmentation cannot be easily applied. This becomes an issue when tabulated datasets are small, which is highly common in medicine and will be covered discussed below.

Another trend in medical data is their typically small size, attributed to the significant time and costs required for data collection and annotation. Additional constraints, such as ethical and privacy concerns surrounding patient information and noises in medical equipment or measurement processes, further limit dataset size. With limited data and especially high-dimensional data, ML models are prone to overfitting, where they overly adapt to the training set and struggle to generalize to new samples [85]. In cases of low data volume, unlike with other modalities (e.g., images, text, and videos), tabulated data in medicine often cannot be easily augmented or synthesized without strong assumptions made [15]. Moreover, performance evaluation on small datasets is inconsistent and unreliable, as it is difficult to determine whether results stem from the models' implementation or random effects arising from a limited test set [98].

Due to the constraints described above, general-purpose DNNs had struggled to achieve State-of-the-Art (SOTA) performance while Gradient Boosted Tree (GBT)s played a dom-

inant role in tabular data classification problems [40, 80, 101]. The term GBT, in this thesis, loosely includes RFs as they are all tree-based and often exhibit similar behavior. However, as more innovative and specialized techniques were continually introduced to deep learning, DNN started to cause a paradigm shift in small tabular datasets, surpassing GBTs' historical dominance [124, 27]. Recently, it has been proposed that the dichotomy of GBTs and DNNs should be discarded, and studies should ensemble both to combine the strengths of different models [27]. This section introduces a series of selected ML models, ranging from instance-based methods, kernel-based methods, GBTs, to eventually DNNs. While methods outside of GBTs and DNNs have been decreasingly popular, they are not completely obsolete in research [107, 63, 120], and can offer representations different from those of GBTs and DNNs.

2.3.1 Instance-Based Methods

2.3.1.1 KNN

KNN is a non-parametric, instance-based learning method, commonly referred to as a lazy learner because it does not learn from the data during the training phase [30]. Instead of learning a discriminant function or the underlying distributions, it operates by simply memorizing all training data and classifying unseen data points by selecting the majority class in the group of K nearest neighbors, as shown in Figure 2.14. KNN is easy to implement, as it requires no training phase and involves only a few hyperparameters to tune, primarily the value of k and the choice of distance metric. This method has been gradually replaced by other techniques in research, but is included for comprehensive coverage of different model classes.

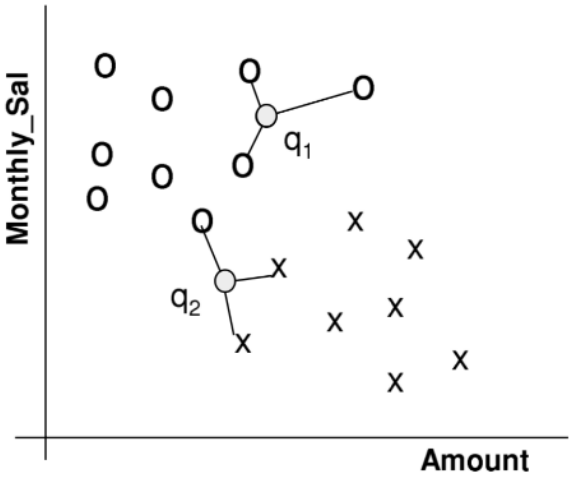


Figure 2.14: An example of 3-nearest neighbor, where q_1 would be classified as "o" and q_2 as "x" [22].

2.3.2 Kernel-Based Methods

2.3.2.1 SVM

SVM is recognized as a class of half-space linear predictors that function within a potentially high-dimensional feature space, the dimensionality of which is determined by the selected kernel [99, 21]. As can be seen in Figure 2.15, SVM operate by establishing a maximum-margin decision boundary within this feature space. The maximization of the margin is intended to enhance the model's generalization ability, reducing sensitivity to noise and outliers in the training data. Common kernel functions, such as the linear, polynomial, and radial basis function kernels, allow SVM to adapt to various data distributions, making them versatile for both linear and nonlinear classification tasks. This operational framework contributes to their robustness and widespread application in ML, particularly for problems requiring clear class separation.

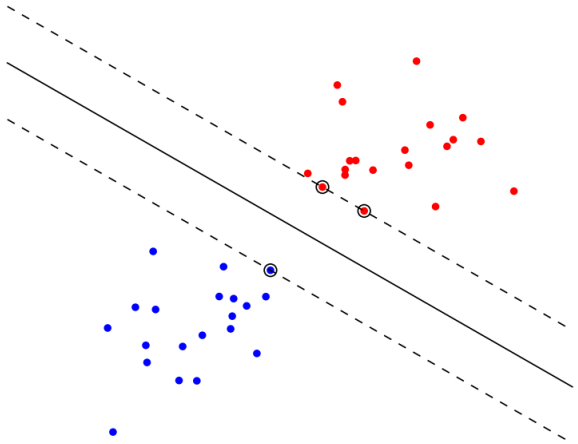


Figure 2.15: An example of linear SVM [72]. Depending on the kernel chosen, the decision boundary can take different forms

2.3.3 GBT

2.3.3.1 RF

RF is an ensemble learning method that that aggregates a large number of decision trees to achieve improved robustness, stability, and generalization performance [99, 17]. Greedy decision trees are defined as a class of tree-like classifiers constructed iteratively from a root node, which initially encompasses the entire dataset. At each node, the data are partitioned into two subsets based on the feature deemed most informative at that point, with each subset forming a new node. This splitting process, depicted in Figure 2.16 is repeated iteratively until a predefined termination criterion is satisfied. Commonly adopted criteria include a maximum number of leaf nodes, maximum tree depth, or threshold for data purity at the leaf nodes, often measured via metrics such as Gini impurity or entropy. Due to their greedy construction and their high expressivity, individual decision trees are likely to overfit. To address this limitation, Random Forests are designed to construct a predetermined number of such trees, each trained on a subset of the data and features, with predictions aggregated through majority voting. This ensemble approach ensures that biases and variances inherent in individual trees are mitigated, thereby improving overall predictive performance.

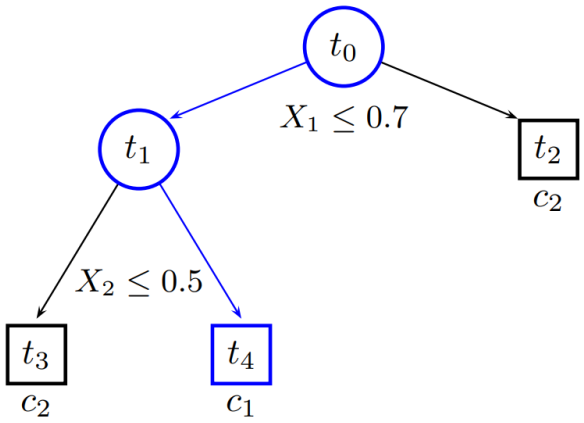


Figure 2.16: Decision tree architecture [72]. A random forest is the ensemble of an arbitrary number of such decision trees, serving as a voting committee

2.3.3.2 XGBoost

XGBoost is acknowledged as an advanced ensemble method that iteratively constructs weak classifiers to enhance the model's performance on a given task [19]. This is achieved by computing the negative gradient of the loss function with respect to the predictions from the previous iteration, followed by the training of a new decision tree to fit this negative gradient to reduce the residual errors. Furthermore, the algorithm's design includes sparsity-aware processing and parallelization capabilities, which are integrated to ensure computational efficiency, scalability, and robustness, even in the presence of missing data. These features enable XGBoost to handle large datasets and complex feature spaces effectively, making it a widely adopted tool for classification and regression tasks in ML.

2.3.3.3 CatBoost

CatBoost, similar to XGBoost, is an ensemble technique based on gradient boosting. However, unlike most gradient boosting algorithms which require categorical features to be converted into numerical representations, CatBoost can handle non-numerical features directly, reducing the time for data pre-processing. Moreover, CatBoost saves time on hyperparameter tuning as it provides high-quality results with default settings.

2.3.3.4 LightGBM

LightGBM is another extension of GBT that was developed by Microsoft [59]. Similar to XGBoost, it builds an ensemble of shallow decision trees sequentially, with each tree correcting the residuals of its predecessors. However, LightGBM introduces several key optimizations that make it particularly efficient. The optimization most relevant for small-to-medium sized datasets is the exclusive feature bundling that reduces feature dimensionality by identifying and merging mutually exclusive features into a single bundled feature, which sparsifies input without loss of information, accelerating and simplifying the search for optimal splits.

2.3.4 DNN

2.3.4.1 MLP

MLP is a DNN composed of stacked perceptrons, represented by Figure 2.17. By the universal approximation theorem, a standard MLP, despite its simplicity, has the ability to approximate any continuous function with an arbitrary accuracy provided it has sufficient hidden units or layers [51, 52]. It is the most simplistic form of a DNN, commonly employed when exploitable inductive biases are unknown or absent.

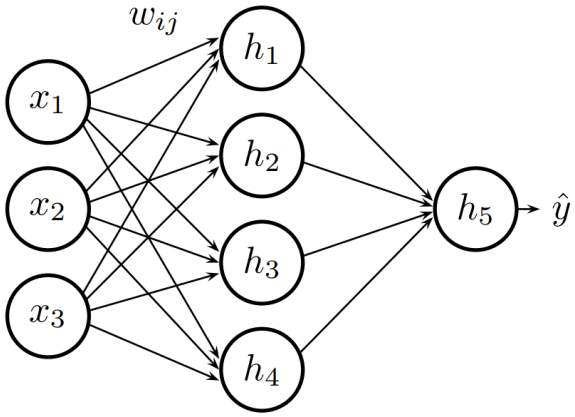


Figure 2.17: MLP architecture [72]. Initially designed to mimic human nervous systems

2.3.5 TabNet

TabNet, proposed in [6], is a DNN architecture designed to unify feature selection, feature extraction, and classification into a single end-to-end framework (Figure 2.18). It consists of two core components shown in Figure 2.19: a feature transformer which learns hierarchical representations of the input features, and an attentive transformer which dynamically selects and weights such representations. These transformers, distinct from the famous self-attention transformer mechanism designed by Vaswani et al. in [111], are built with Fully Connected Layers (FC)-layers, batch normalization, Gated Linear Units, Sparsemax, and learnable *prior scales* that encourage feature sparsity and interpretability. Due to the dominance of GBTs in tabular data, TabNet incorporated result aggregation for GBT-like behavior. It performs reasoning sequentially, transforming data and attending to them differently at each decision step. Part of the presentations at each decision step gets passed along to the subsequent decision step, and the local output at each decision step are aggregated to yield a final global output.

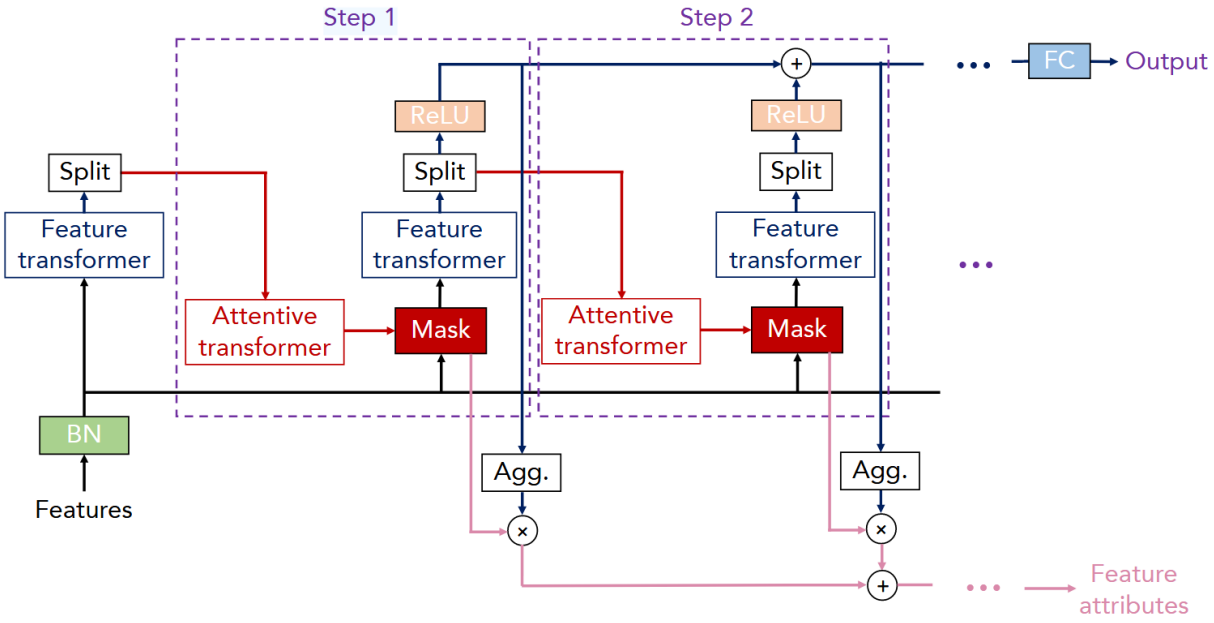


Figure 2.18: TabNet Architecture, designed by Google [6]. The Upper stream yields predictions while the bottom stream provides feature importance

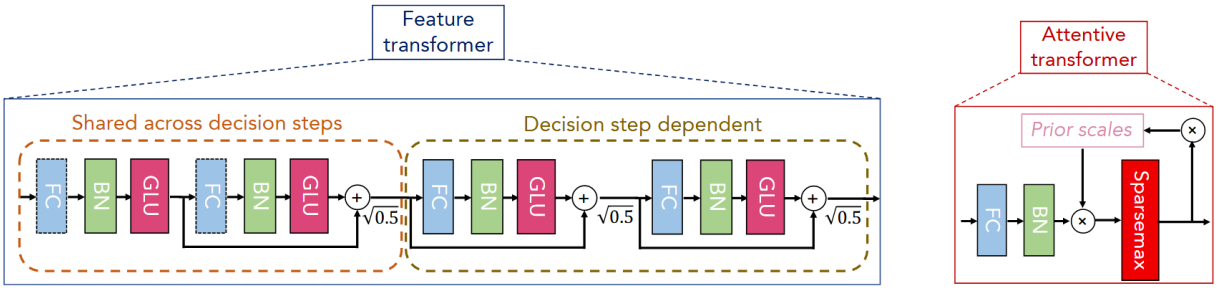


Figure 2.19: TabNet feature and attentive tranformers [6]. Despite the name *transformer*, they are unrelated to the famous self-attention-based transfomer

Additionally, TabNet provides a self-supervised pre-training learning paradigm, in which a subset of input features is randomly masked, and the model is trained to reconstruct the masked values. This process implicitly coerces the model to capture the inherent structure within the data without relying on labels. The pre-trained weights, encoding knowledge of feature dependencies and distributions, function as an initialized backbone for the supervised TabNet classifier during fine-tuning. This transfer learning requires an encoder-decoder structure as shown in Figure 2.20.

2.3.5.1 TabM

TabM is an ensembled extension of the MLP. It was proposed in [37] to overcome the computational inefficiency of traditional deep ensembles which require training multiple independent models from scratch. TabM leverages batch ensembles, a parameter-efficient ensembling technique that enables training multiple ensemble members in a single forward-backward pass. As depicted in Figure 2.21, each layer in TabM contains a shared base weight matrix along with per-member perturbations that can simultaneously be trained or perform inference. This design yields an ensemble of diverse predictors at a fraction of the cost of deep ensembles, while providing uncertainty estimates similar to those of Bayesian neural networks via the predictive distribution over members.

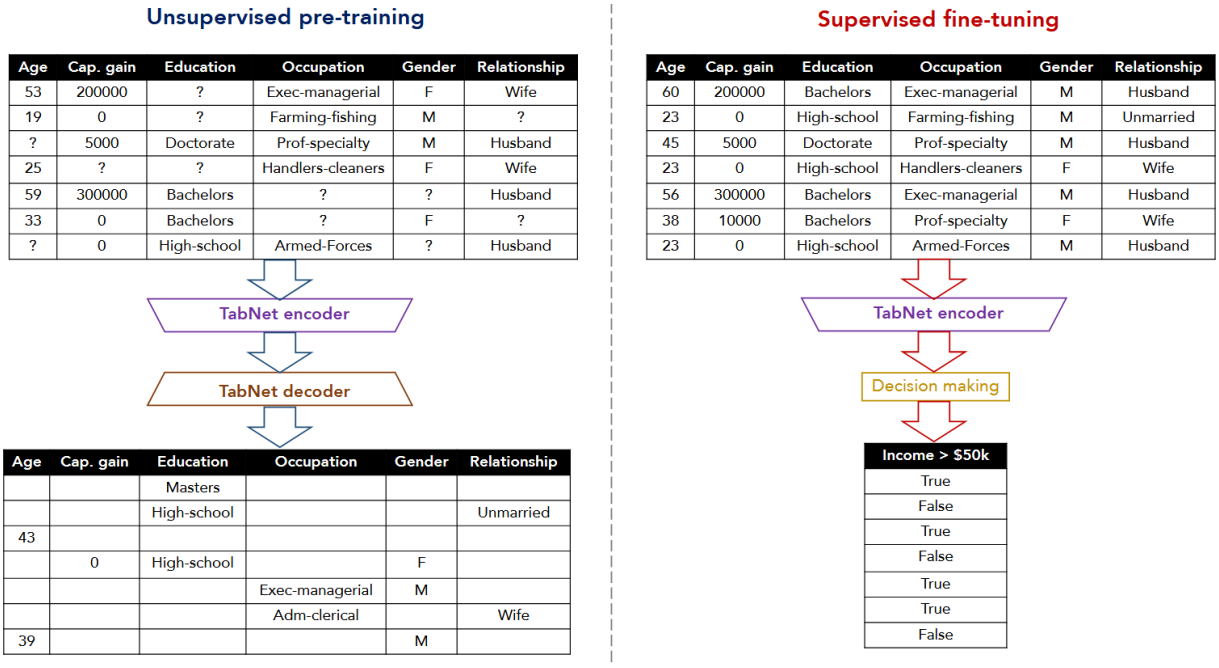


Figure 2.20: TabNet encoder-decoder structure for self-supervised learning [6]

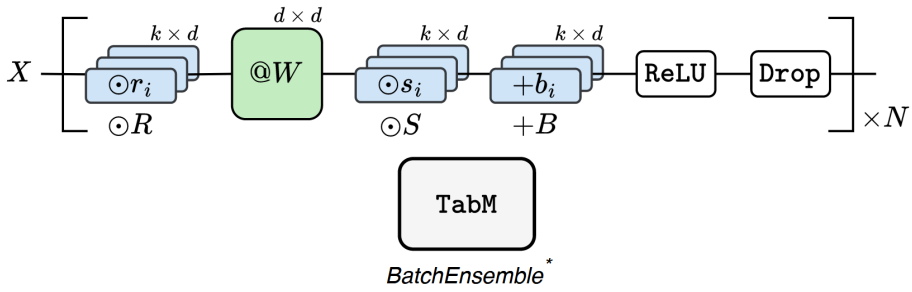


Figure 2.21: TabM architecture [37]

2.3.5.2 TabPFN v2

TabPFN is an In-Context Learning (ICL) method that eliminates traditional training, hyperparameter-tuning, and inference, replacing them with a forward pass in a pre-trained Transformer model [50]. It is an end-to-end Prior-Data Fitted Network (PFN) that approximates the Bayesian posterior inference process over functions by implicitly learning during pre-training the proper hypotheses that describes different data generating mech-

anisms. Once it has learned the inherent mappings between inputs and outputs in all its train datasets from the pre-training phase, it can be conditioned on a subset of the dataset of interest and predict labels for unseen data entries.

Its later version, TabPFN v2 [121], then introduces (I) extra embedding layers and positional encoding for explicit 2D self-attention (cross-sample and cross-feature) and (II) random perturbation to features and removes feature names for automatic feature identification, achieving effects similar to those of Rotary Positional Encoding (RoPE). This constrained the TabPFN v2 to learn solely on the feature content rather than being able to rely on feature names or feature IDs for easier solutions.

This is a new paradigm that challenge the gold standard of ML practice, reported outperforming the traditional learning paradigm. An image description of TabPFN v2 is shown in Figure 2.22.

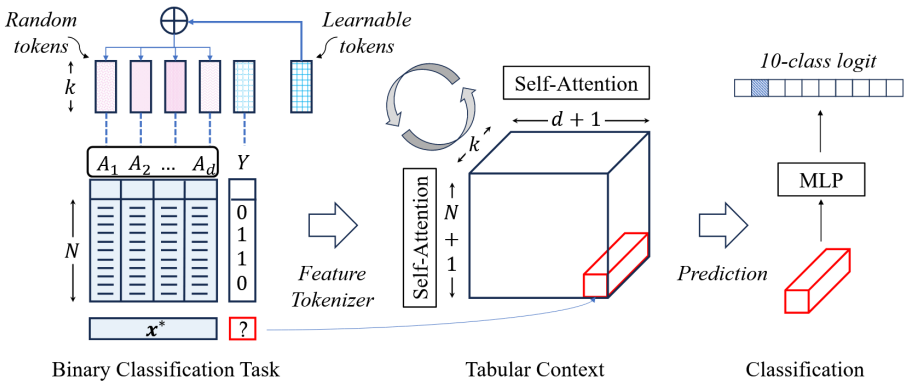


Figure 2.22: TabPFN architecture [121]. This architecture is different from TabPFN v1

2.3.5.3 TabICL

TabICL is a tabular foundation model inspired by TabPFN, adopting the same ICL paradigm to bypass traditional training and inference loops [91]. Similar to TabPFN, it is pre-trained on a massive corpus of synthetic tabular datasets to learn a prior over underlying input–output mappings, enabling prediction on unseen data by conditioning on a small labeled context in a single forward pass. The main structural difference between the two models is that TabICL divides the two-way attention mechanism into a two-phase set transformer attention mechanism (Figures 2.23 and 2.24) and an ICL transformer (Figure 2.25), compressing inputs into more dense representations for more efficient computation. Additionally, TabICL replaces the random perturbations introduced in TabPFN v2 by

RoPE. These modifications make TabICl more stable and generalizable to in-context predictions.

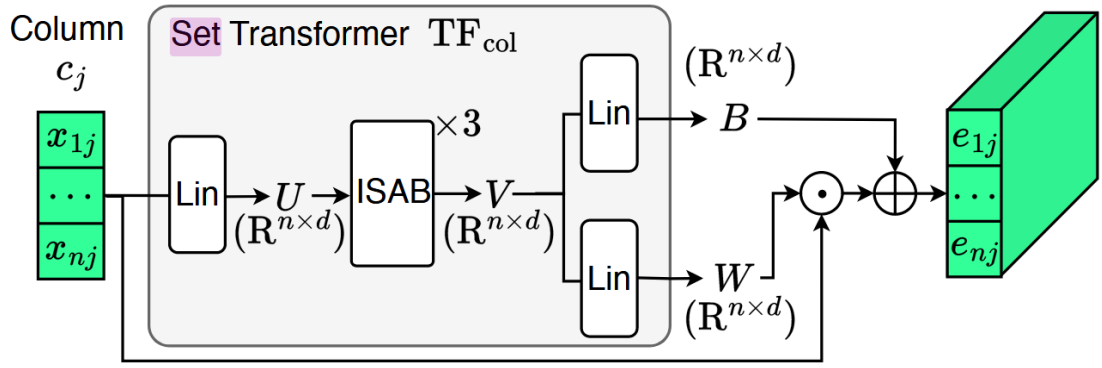


Figure 2.23: TabICL TF_{col} architecture [91]

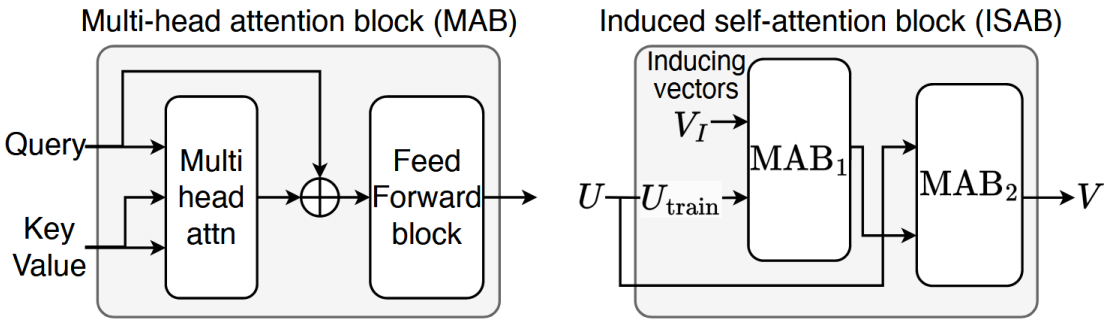


Figure 2.24: TabICL TF_{col} sub-architecture [91]

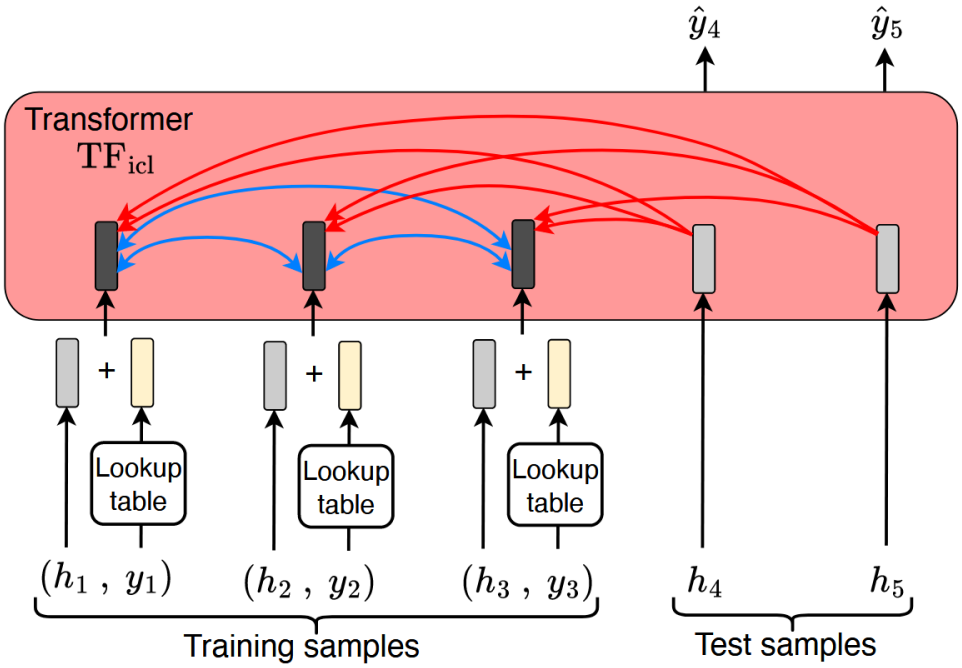


Figure 2.25: TabICL TF_{icl} architecture [91]

2.3.5.4 MITRA

MITRA, presented in [127], is a model that highly resembles TabICL. Despite being developed by Amazon Web Service and demonstrating superior performance according to [127], the implementation details are less documented. It is included as it is one of the recommended options supported by AutoGluon.

2.4 Gaps in Current Research

Despite the growing recognition of tibial slopes and depths as critical predictors of ACL injury risk, few ML studies adopted explainable SOTA ML models to examine tibial anatomical measurements, highlighting a research gap. A study by Hashemi et al. developed a logistic regression model that identified MTS, LTS, and MTD as the most important risk factors in ACL injury prognosis [41]. Kiapou et al. constructed bivariate linear regression and multivariate stepwise regression models which established MTD and slopes in the coronal and sagittal planes as risk factors [61]. Both studies selected linear models that are

not guaranteed to represent the true relationship between TAFs and ACL injury outcomes, and focused on analysing correlation rather than predicting risk of injury. Tamimi et al. employed ML and knee anatomic features such as various tibial slopes and meniscal heights to predict ACL injury with 70% validation accuracy and 92% testing accuracy [108]. The primary limitation of this study, though, was the use of Gaussian Naïve Bayes models, which assumed conditional independence between features, potentially overlooking interactions among tibial anatomical measurements that could enhance predictive performance. Due to the lack of cross-feature interaction terms in the Gaussian Naïve Bayes formulation, any joint effects among features cannot be modelled unless interaction terms are manually integrated, a process that is neither exhaustive nor scalable.

ACLNet [70], a recent multi-modal model developed by Liu et al., leveraged SOTA model architectures to make high-accuracy predictions based on MRI images and bone morphological point clouds. Although this model is designed primarily for diagnosis rather than prognosis, it is included here as it represents the emerging class of models that utilize unstructured data to predict injury. These approaches, despite their high performance, inherently lack interpretability because the model relies on complex, unstructured input data rather than specific, medically relevant features. The absence of a clear link between a high-injury-likelihood prediction and a measurable anatomical feature prevents clinicians from justifying an intervention, which is a major limitation.

This thesis highlights addresses a gap in existing literature, which is the absence of explainable and SOTA ML models in studying TAFs to elucidate the most relevant predictors from among them. The goal of the current study is to fill this gap by carefully selecting SOTA ML models designed to handle small tabular data, experiment with different variants and mechanisms, and further utilize feature importance analyses to identify the most contributive tibial anatomical variables.

Chapter 3

Methodology

3.1 Dataset

The same dataset is used by the two pipelines, presented to both in the exact same format and order.

3.1.1 Source

The dataset used in the study was originally collected by Hashemi et al. [41, 42]. It consisted of TAFs such as CTS, MTS, LTS, and MTD that were manually extracted from MRI scans from 49 ACL-injured and 55 uninjured participants. The detailed mean and standard deviation statistics for the different groups are presented in Table 3.1, and the raw dataset is provided in A.

Table 3.1: Mean and standard deviation statistics for the injured and uninjured groups of male and female participants, computed from the entire dataset. They are expressed in the format of *mean* $\pm$ *standard deviation*

Measure	**Uninjured Male**	**Injured Male**	**Uninjured Female**	**Injured Female**
Group Size	22	22	33	27
CTS (°)	3.50 ± 1.87	3.27 ± 2.53	2.61 ± 1.87	2.59 ± 2.12
MTS (°)	3.68 ± 3.11	6.09 ± 2.71	5.94 ± 2.95	6.89 ± 3.61
LTS (°)	5.41 ± 2.77	7.23 ± 2.71	7.12 ± 2.91	8.44 ± 2.82
MTD (mm)	3.05 ± 0.98	1.98 ± 0.98	2.74 ± 0.76	1.91 ± 0.98

Among the data collected, CTS, MTS, LTS, MTD, and sex were designated as input features, while injury served as the target label. The features comprised both quantitative variables (CTS, MTS, LTS, and MTD) and a categorical variable (sex), whereas the injury label was also categorical. More specifically, the tibial slope measurements (CTS, MTS, and LTS) were recorded as integers, and the MTD values were represented as floating-point numbers with two decimal places of precision. All categorical feature were one-hot encoded, resulting in a preprocessed dataset containing 104 instances, five features, and two label classes.

Although Principal Component Analysis (PCA) is commonly employed for dimensionality reduction and feature transformation, it was not considered in this study due to two key limitations. First, the embeddings produced by PCA are inherently linear [56], incapable of capturing any underlying non-linear interactions between features. Second, despite its linear nature, the feature importance attributed to PCA components cannot be easily mapped back to the feature space, which reduces the interpretability and clinical relevance of the prognostic model developed.

UMAP, as another dimensionality reduction tool, shares the same limitation in interpretability. Nevertheless, it functions as a practical tool for data visualization in the lower-dimensional space, as it preserves both global and local structure in the data [81]. In contrast, t-SNE was not selected for dimensionality reduction because it preserves only local structure, often distorts global geometry, and is more sensitive to hyperparameters [118].

Accordingly, the original features set was projected via UMAP embeddings to a 2D space in Figure 3.1 and a 3D space Figure 3.2 for visualization. It is evident from these projections that there are no exploitable structures in the raw feature set.

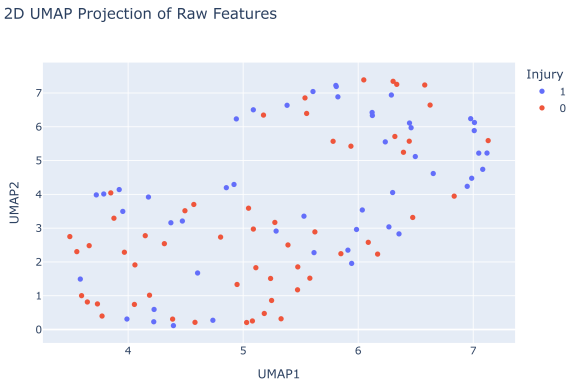


Figure 3.1: 2D UMAP projection of raw data

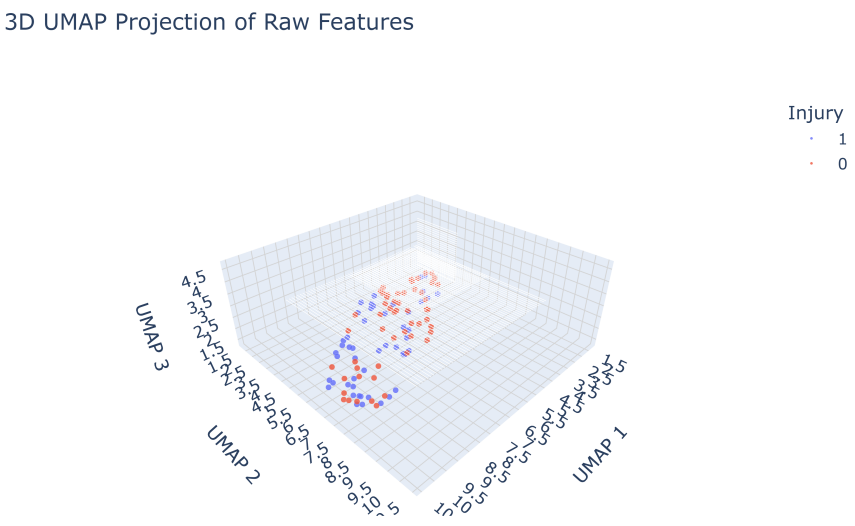


Figure 3.2: 3D UMAP projection of raw data

3.1.2 Feature Selection

The four TAFs were selected and extracted from the MRI scans by Hashemi et al. [42] based on experts' domain knowledge, each containing clinically relevant senmantics. Considering all TAFs and sex were features of interest, and that no significant co-linearity was detected, no feature pruning was required. Feature selection was then handled by the two pipelines differently. The self-developed pipeline applied Recursive Feature Elimination (RFE) and exhaustively evaluated all possible feature combinations across all ML models, except for TabNet, which explicitly performs feature selection as part of its architecture. In contrast, the AutoGluon pipeline relied on the *AutoMLPipelineFeatureGenerator* for additional feature engineering (including data-type inference, missing-value handling, and data pruning) and performed implicit feature selection through its multi-layer stack ensemble.

3.1.3 Data Synthesis

Since the dataset was roughly balanced (47.12% injured and 52.88 % uninjured), imbalanced-learning techniques were deemed unnecessary. Additionally, data augmentation and synthesis are unsuitable due to constraints imposed by the studied dataset. Unlike structured modalities such as text or images, the tabular data containing tibial anatomical parameters lack inherent spatial or sequential structure. The absence of inductive biases renders data augmentation and synthetic data generation substantially more challenging [112]. Similarly, other methods such as oversampling, statistical perturbation, Generative Adversarial Networks, and Variationl Autoencoders pose additional risks in the absence of sufficient domain knowledge, particularly given the clinical nature of the injury-related prediction task [62]. This introduces a methodological paradox, where reliable data synthesis would require established domain knowledge, yet acquiring such knowledge is precisely one of the primary objectives of this study. Hence, applying such techniques prematurely would likely reinforce unknown biases rather than mitigate them. Givens the limitations of data synthesis methods described above, this study relies solely on the data available without introducing any inductive artifacts, despite the small dataset. The implication of this design choice will be discussed in Chapter Chapter 7.

3.1.4 Metric Selection

Despite the widespread of use of of accuracy, F1-score, and AUC [3], recent reviews emphasize that no single metric is universally optimal and that the choice should reflect both dataset characteristics and intended research questions [95, 49]. Given that missing a

true positive is clinically more detrimental than incorrectly flagging a negative case in the context of ACL injury prognosis, the main objective of this study is to minimize false negatives, even at the cost of a moderately elevated false-positive rate. The F2-score, which favors recall/sensitivity over precision, was consequently used as the primary metric for hyperparameter tuning in this thesis. For completeness and to offer clinicians greater flexibility, models were also tuned with respect to accuracy and F1-score, and will be provided as supplementary alternatives. A more comprehensive set of metrics are presented during evaluation, as covered in Chapter 4.

3.2 Self-Developed Pipeline

At the time implementing the self-developed pipeline, most research evidence pointed towards the superiority of GBTs for small tabular datasets [101, 40, 80]. As a result, the pipeline was primarily designed around SOTA GBT implementations, while a small number of models from other classifier families were also included for completeness. The specific model chosen for this pipeline can be found in Figure 3.3.

This pipeline was enabled by Python 3.12.4, integrated with models built using TabNet 4.1.0, CatBoost 1.2.7, Scikit-Learn 1.5.1 [87], XGBoost 2.1.1, and PyTorch 2.4.0. CUDA 12.6.0 was enabled for GPU parallelized acceleration.

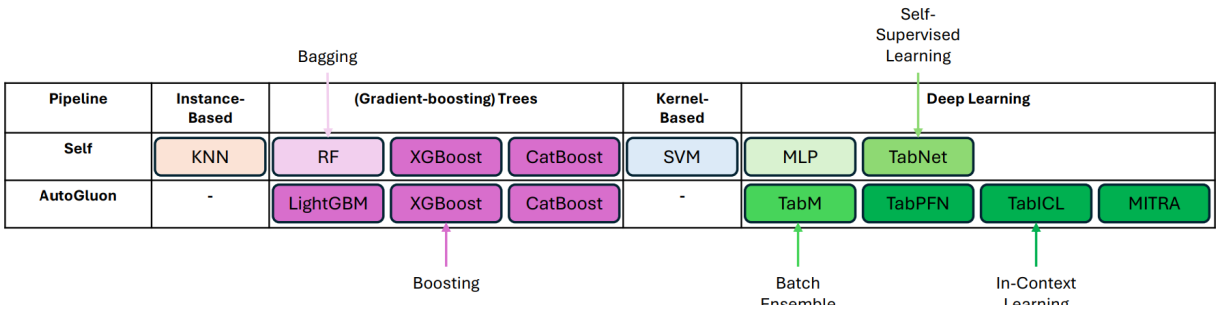


Figure 3.3: Learning paradigm comparison between two pipelines

3.2.1 Training and Tuning

Training served as the core inner loop of the pipeline and adheres to standard practices. For most models included in this pipeline, default library configurations were used for training with two exception: (I) all random seeds are fixed to 42 for fairness and reproducibility,

and (II) mini-batch training was disabled/avoided and full-batch updates were used due to constraints posed by the use of time-invariant features for time-dependent outcome predictions, as will be discussed in Chapter 7. The termination condition for each model varied, depending on their specific implementation. All DNN models were allowed to train for a maximum of 10000 epochs, but with early stopping enabled. All GBTs' termination conditions were dictated by the maximum number of splits, depth, leaves, and samples. XGBoost and CatBoost additionally introduced early stopping. Since KNN does not have a learning process and is fully instance-based, no termination condition was required. SVM, following the default implementation, dictated the termination by monitoring the optimization process through convergence tolerance and maximum number of iterations. All models were trained on the CPU due to empirical observations of CPU-training running faster than GPU-training, but had CUDA enabled (for those that support CUDA) for potential GPU-training. The speed difference between CPU- and CUDA-based training is due to the small size of the dataset, which is latency- and overhead-bound rather than compute- or memory-bound [39].

Due to preliminary observation of overwhelming overfitting (Figure 3.4) in early experiments, the architecture of the conventional MLP in this study was enhanced with skip connections, following the residual learning framework of [44], to improve more training stability and convergence speed. Each layer block in the model therefore comprises an FC-layer, an activation layer, a dropout layer, a residual connection layer, and a batch normalization layer. The full network concludes with one final FC-layer and a softmax layer, a configuration used to normalize logit vectors to probability vectors. To further mitigate the risk of overfitting, model weights were penalized with Elastic Net regularization, which combines the sparsity-inducing L1 norm with the coefficient-shrinking L2 norm to maintain a balance between feature selection and stability. Additionally, early stopping with a moving-average monitor was implemented. After a short warm-up phase of *warm-up* epochs, validation loss is tracked over a sliding window of length *moving_average*. If the latest validation loss exceeds this moving average for *patience* consecutive epochs, training terminates and the model reverts to the parameters recorded at the epoch when the loss first crossed the average such that the saved checkpoint corresponds to peak generalization performance.

No dedicated training/validation split was used. Instead, 10-fold cross-validation was employed to maximize data utilization, partitioning the combined train-validation set into ten non-overlapping folds. In general, k-fold cross-validation splits the train-validation set into k splits, each containing k-1 folds of test subsets and 1 fold of validation subset (Figure 3.5). During hyperparameter tuning, the same configuration is applied across all splits. The average validation performance across all splits serves as a robust, low-bias

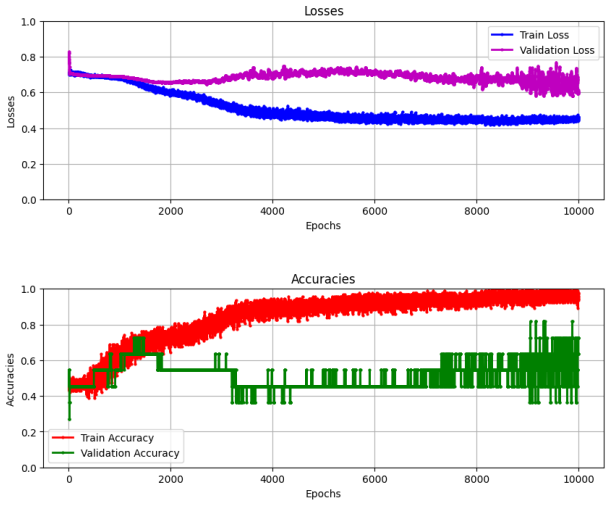


Figure 3.4: Preliminary results manifesting overfitting

estimate of generalization, and is used for guiding the hyperparameter tuning process. This, while more data-efficient and robust, complicates early stopping. While models from all splits share identical hyperparameters, the optimal number of epochs varies depending on training/validation the specific split. This is because hyperparameters *patience* and *warm-up* describe the early stopping dynamics but do not explicitly specify the number of epochs used during training. Since the final model must be trained on the entire train-validation set and cannot leverage an independent held-out set for determining the optimal number of epochs (as that would defeat the purpose of k-fold cross-validation), the pipeline took the average of the optimal number of epochs as an unbiased estimate, and trained the final model accordingly. While this substantially improved data usage, it also increased the computational overhead by a factor of k which was ten times in this case.

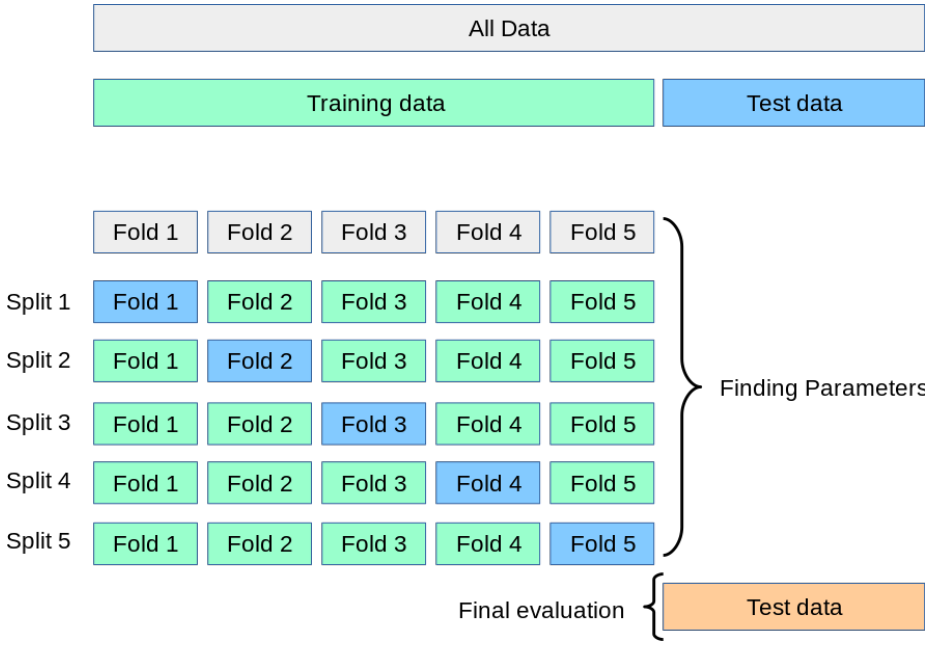


Figure 3.5: An example of 5-fold cross-validation [87]

Furthermore, the homogeneous bagging mechanism proposed in [16] was enabled as an option in the self-developed pipeline. Homogeneous bagging creates multiple instances of a base learner, each trained on a unique bootstrap sample drawn with replacement from the full train-validation set. Each bootstrap sample matches the original dataset's size, typically containing approximately 63% unique data points, with the remainder being duplicates. These bag models are tuned and trained in parallel, sharing hyperparameters but fitting independently to their respective bootstrap samples. Final predictions from each bag model are aggregated via majority voting for classification, yielding a robust output less prone to overfitting. By leveraging diversity among models trained on varied folds, homogeneous bagging reduces variance without significantly increasing bias. The entire bagged ensemble is treated as a super learner and undergoes full hyperparameter tuning. Although fully supported, bagging was excluded from the scope of the thesis due to its substantial computational overhead (2-10× increase in training time). Preliminary experimentation results in earlier stages of the thesis, tabulated in Table 3.2, have shown some evidence that homogeneous bagging moderately improves generalization on the dataset.

Table 3.2: Change in validation and test performance when homogeneous bagging is applied

Model	Validation		Test	
	Accuracy	F1 Score	Accuracy	F1 Score
MLP	0.0312	0.2410	-0.0909	0.2122
SVM	-0.0097	0.0045	0.0909	0.0556
KNN	0.0000	-0.0101	-0.0909	-0.1389
RF	-0.0313	-0.0432	0.0909	0.1389
XGBoost	0.0132	0.0404	0.0909	0.1667
CatBoost	-0.0208	0.0078	0.0909	0.2143

The hyperparameter tuning process was realized through Bayesian optimization, more specifically Optuna 4.4.0 [2]. It is model-agnostic, requiring no assumptions about the loss surface [32]. It is also significantly more sample-efficient than other naive approaches like grid search or random search, as it more intelligently searches through the hyperparameter space with Tree-structured Parzen Estimators. A budget of 10000 trials is allocated per model to ensure thorough exploration of the search space, with an example presented in Figure 3.6.

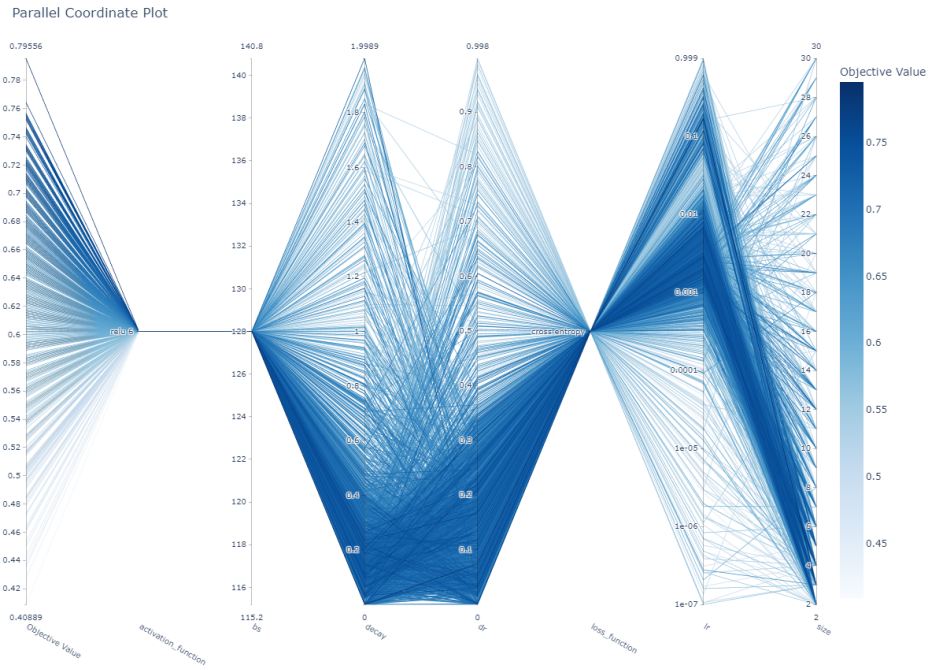


Figure 3.6: Bayesian optimization history

3.2.2 Evaluation

The models were selected based on the target metric used in hyperparameter-tuning. The best-performing model corresponding to each tuning strategy was then evaluated on the held-out test set. Their validation and test performance were both tabulated and reported.

3.3 AutoGluon Pipeline

The experiments conducted in the AutoGluon pipeline were carried out within a Python 3.11.13 environment, leveraging the ML capabilities provided by AutoGluon 1.4.0. AutoGluon-Tabular (referred to as AutoGluon for conciseness) is an AutoML framework designed by Erikson et al. to handle tabular data tasks [26]. Its *extreme* preset, inspired by TabArena [27], was recently added in the latest version 1.4.0. to tackle small datasets. By strategically combining GBTs, DNNs, and ensembling them effectively, AutoGluon was able to achieve SOTA performance on small tabular predictions problems. Consequently, the ML pipeline adopted in this section was built upon AutoGluon, capitalizing on its highly optimized and robust implementations to predict ACL injury.

Following work like [1], AutoGluon recognizes the complementary nature of GBTs and DNNs rather than simply viewing them as competing techniques with superset relationships. It explicitly exploits model diversity across classifier families by combining them to construct powerful stack ensembles, thereby improving robustness and generalization. The specific model chosen for this pipeline can be found in Figure 3.3.

3.3.1 Training and Tuning

The AutoGluon pipeline adopted a training and ensembling strategy that closely mirrors that of the self-developed pipeline, involving both bagging and k-fold cross-validation. However, the former integrated them in a more sophisticated and efficient manner via k-fold bagging, a technique innovated by Parmanto et al. in [84]. Rather than applying bagging and k-fold cross validation sequentially or independently, k-fold bagging unifies them by preserving the k fold models obtained from each of the k splits and ensembling them into a single bag model. Identical to k-fold cross validation, each of the fold models in the bag model is unique since they are trained on partially overlapping but different folds of data. By forcing diversity in the fold models, this approach obtains a bag model of high expressivity in nature. A schematic of the workflow is shown in Figure 3.7.

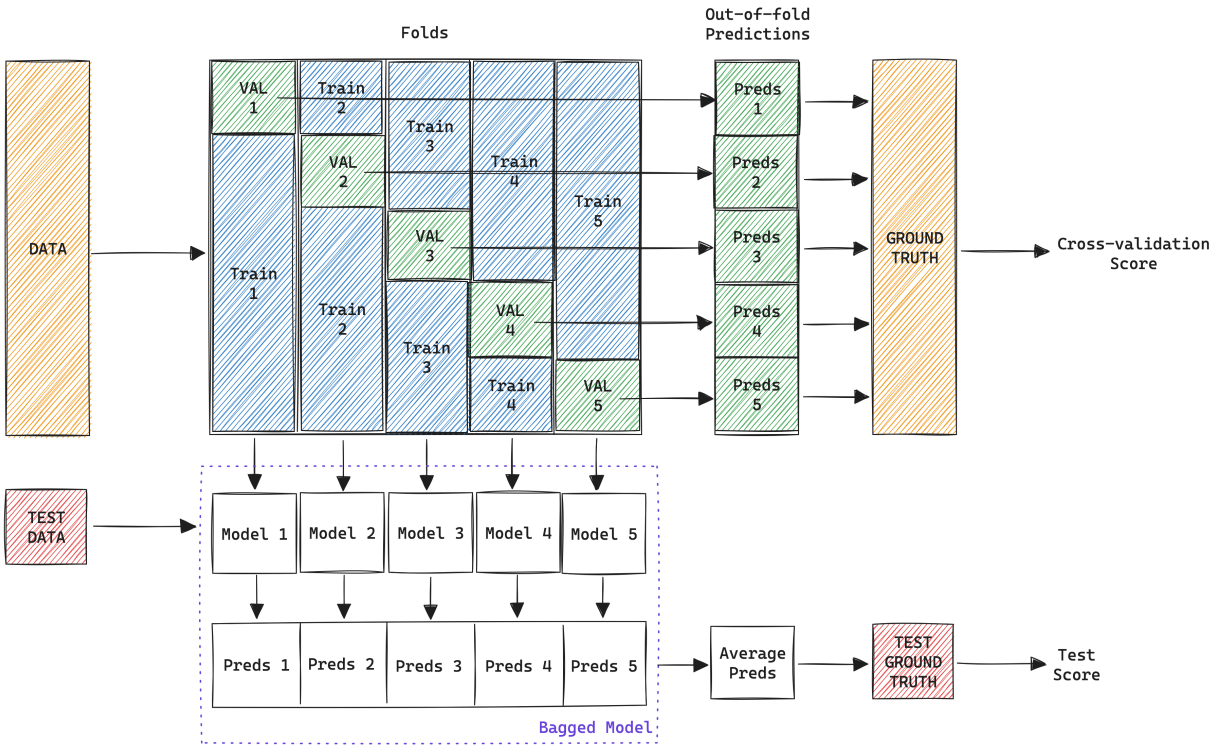


Figure 3.7: K-fold bagging [26]

This design was further enhanced by multi-layer stacking proposed in [26]. This was analogous to layer-wise training of DNNs, where each *neuron* (in this case a bag model) learns a highly nonlinear representation that gets passed to the *neurons* in the proceeding layer. Similar to the MLP implemented in the self-developed pipeline and many existing DNNs, the input features are passed to deeper layers through *skip connections*, ensuring both shallow-layer representations and intermediate-layer representations are accessible to deeper layers. The last layer is similar to the FC-layer in most DNNs, performing linear weighted sum to aggregate representations learned in the last *hidden layer*.

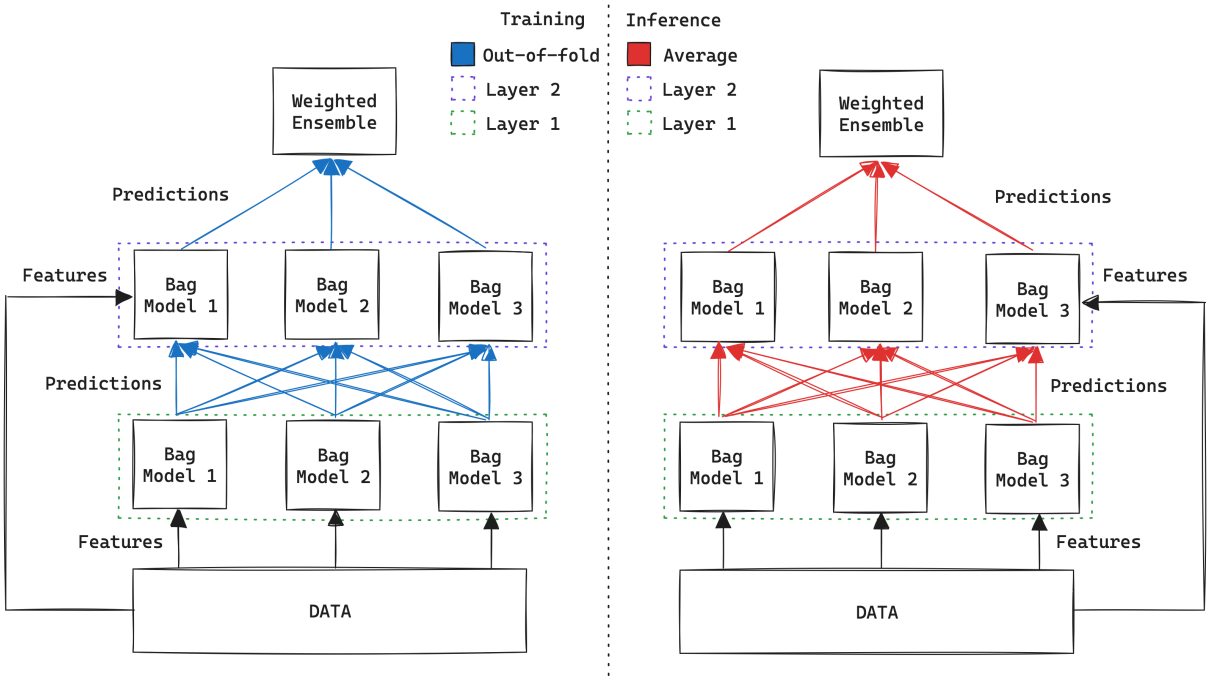


Figure 3.8: Multi-layer stacking [26]

Avoiding data leakage during training is slightly more convoluted in multi-layer stack ensemble, and is the major reason of adopting k-fold bagging in this workflow. For each bag model, since the held-out set from each of the fold models are non-overlapping, each instance in the train-val set can be used to generate an out-of-fold prediction by a fold model that was not trained on that instance. More specifically, any data point in a fold *VAL* i from Figure 3.7 should be passed to the corresponding fold model *Model* i for Out-of-Fold (OOF) prediction. It should be noted that the validation folds drawn in the figure, unlike the standard k-fold cross-validation scheme, is not used for hyperparameter tuning or early stopping. Rather, its sole purpose is to provide OOF predictions such that data leakage is avoided. Data for hyperparameter tuning is extracted from the train folds, which is internally handled by AutoGluon. Iterating through the entire feature set and finding matching fold models yields a complete set of OOF predictions. These OOF predictions are crucial for training deeper layers in the multi-layer stack ensemble, as they are unbiased with respect to the original training labels while maintaining data integrity. They are passed to deeper layer in concatenation with original features, enabling the aforementioned *skip connection* for information flow from both the input and intermediate layers while allowing independent control. AutoGluon's learning scheme is performed iteratively and

sequentially. Bag models in a shallower layer are trained before bag models in the next are trained. This greedy approach forces model characteristics to be finalized before moving to the next level.

According to [26], manual specification of custom hyperparameters is strongly discouraged due to AutoGluon's highly optimized, intelligent, and adaptive internal tuning mechanisms. The default strategies, previously refined across diverse benchmarks and real-world datasets, systematically explores through hyperparameter spaces, enforces model-specific best practices, and adapts searching based on data characteristics. As a result, all hyperparametes and their search ranges were set to the default values provided by AutoGluon.

3.3.2 Evaluation

Leveraging AutoGluon's exceptional computational efficiency, experiments could be executed with reduced runtime and overhead. This enabled a more rigorous and statistically robust evaluation of the AutoGluon pipeline compared to the self-developed pipeline. While the core evaluation protocol remains nearly identical, the AutoGluon pipeline was executed across ten independent random seed configurations to quantify performance stability. In addition to the default seed of 42, nine additional seeds were evaluated, each incremented by ten. For each seed, the complete AutoGluon workflow was executed, including data splitting and pre-processing, training of bag models and multi-layer stack ensembles, internal hyperparameter tuning, and evaluation on respective held-out test sets. All performance metrics were averaged across all seeds to determine the best-performing model type. The instance of the best-performing model from the default seed of 42 (same as in the self-developed pipeline) was selected as the best-performing model.

3.4 Additional Analyses

Obtaining predictive models are necessary but insufficient for clinical adoption. This part of the study investigates key properties of the trained models beyond raw performance, including their stability, discriminability, explainability, and suitability for deployment. These analyses bridge the gap between statistical performance and real-world clinical utility, ensuring the selected models are not only accurate but also reliable, transparent, and practical in diagnostic settings.

3.4.1 Prediction Confidence and Feature Importance Explanation

To enhance system transparency and support clinical interpretability, all predictions were accompanied with class probability estimates. These probabilities, accessible through *predict_proba()* and other similar methods serve two purposes: (I) clinical decision supported by quantifying prediction confidence, and (II) enabling post-hoc FIA via mathematical techniques like SHAP.

A trade-off frequently exists between model performance and explainability, as more expressive models often exhibit greater complexity. To interpret black-box ML models in biomechanical studies, SHAP [77] has emerged as the gold standard, commonly used for feature selection, importance evaluation, and model explanation [71, 55, 122]. It has gained prominence in ML due to its capacity to quantify each feature's contribution to a prediction [77]. Some studies further extend this framework to analyse the contribution of individual components within model architectures [36, 4]. This method, grounded in cooperative game theory, considers all possible coalitions and compares the change in model output to determine the fair contribution of each feature. SHAP values for each feature is guaranteed to follow four axioms: efficiency, symmetry, dummy, and additivity, which makes it theoretically more sound in comparison to heuristic alternatives like LIME or naive feature permutation. In this study, SHAP 0.46.0 was integrated into both pipeline to compute SHAP values for each feature.

More specifically, the SHAP explainer used in this study was the Kernel SHAP. It is a model agnostic explainer applied to all models for fair comparisons. SHAP explainers quantified the contribution of each TAF to the model's predictive output at two levels. The first was a local explanation which computes the amount by which each TAF positively or negatively affect the predictions of an individual with respect to the prediction baseline. This serves as more granular and personalized analysis ideal for individual judgment and potential interference. The second was a global explanation computed by aggregating the local SHAP values among patients, which provides more general patterns useful for trend identification and reletive feature importance quantification. These explanations were visualized in easy-to-read plots, which allow physicians to understand the rationale behind the model's high- or low-injury-likelihood predictions of ACL injury. This adds trust and makes the tool more usable in clinical decision-making.

3.4.2 Full-dataset Model

Leveraging insights from prior sections and acknowledging the high overfitting risk in small datasets, the final models were retrained on all available data to maximize predictive power and generalization.

AutoGluon provides a native *refit_full()* method that reuses the optimal ensemble architecture and hyperparameters identified during cross-validation and retrains the selected models on the full train-validation set. This mirrors the custom pipeline's final retraining step, in which the optimal number of epochs/rounds averaged across k folds and best hyperparameters were applied without early stopping or a held-out validation set. The *refit_full()* approach, known as model collapsing, replaces the k fold models with single, full-data instances. This reduces computational complexity caused by k-fold bagging as only one instance has to be trained per original bag model, as opposed to k instances. While computationally appealing, this sacrifices bagging-induced variance reduction and can degrade performance, especially in small-dataset use cases where ensemble diversity is critical.

Given that inference speed and memory constraints are not primary concerns in the target clinical deployment scenario, and in order to avoid reliance on potentially unrepresentative train-validation splits, *refit_full()* was deemed suboptimal. Instead, a full-dataset model was trained on the entirety of the original dataset (train + validation + test), treating it as a unified train-validation pool. This method still allows AutoGluon to tune hyperparameters and construct k-fold baggings with OOF predictions, such that the risk of overfitting to the new test folds is minimized. This strategy statistically yields the most generalizable and high-performing model possible from the data, ideal for deployment in clinical practice. However, it consumes the independent test set, making unbiased performance evaluation unfeasible. Therefore, all rigorous analyses, comparisons, and conclusions were based exclusively on the original standard ensemble model and the original test set from the standard train/validation/test protocol. The full-dataset model was designated solely for deployment, not scientific reporting.

Chapter 4

Results

4.1 Self-Developed Pipeline

The self-developed pipeline is computationally expensive on account of RFE, large number of Bayesian optimization trials, and lack of systematic pipeline optimization (like AutoGluon does). Due to time constraints, only two tuning metrics, balanced accuracy and F1-score, are used for hyperparameter tuning.

4.1.1 Test Performance

The best-performing model from each of the model families are recorded and tabulated in Tables 4.1 and 4.3. With the exception of TabNet, all models were trained and validated using RFE, systematically testing all non-empty subsets of the five input features. With 31 RFE variants per model (except TabNet) per hyperparameter tuning metric, each table entry corresponding to a tuning metric represents the top-performing configuration selected from 187 different model instances.

The confusion matrices corresponding to the best-average-performing models are presented in Figure 4.1 and 4.2 for visualization of predictive performance and transparency, as it can be used to derive other unreported metrics.

4.1.1.1 Accuracy-Tuned TabNet

The best-performing accuracy-tuned models from each model family are included in Table 4.1. TabNet, despite relying solely the full feature set, achieves the highest validation

performance across all model types, slightly beating the second candidate model by 0.032. It generalizes well, suffering a loss of 0.050 in accuracy when performing inference on the test set. This TabNet model, referred to as the *accuracy-tuned TabNet*, has the finalized hyperparameters as summarized and described in Table 4.2.

Table 4.1: Best-performing models with the highest validation accuracies from each model family and their corresponding feature subsets. The integers 1 to 5 in the table represent different features - 1:CTS, 2: MTS, 3: LTS, 4:MTD, 5:sex

Model	**Features**	**Val Acc**	**Test Acc**	**Test F1**	**Test F2**	**Test AUC**
TabNet	[1, 2, 3, 4, 5]	0.863	0.812	0.800	0.769	0.959
CatBoost	[1, 2, 3, 4, 5]	0.831	0.812	0.769	0.676	0.890
XGBoost	[3, 4]	0.830	1.000	1.000	1.000	0.961
RF	[1, 3, 4]	0.795	0.750	0.667	0.556	0.896
SVM	[1, 2, 4, 5]	0.761	0.812	0.824	0.854	0.840
KNN	[4, 5]	0.750	0.875	0.857	0.789	0.874
MLP	[2, 3]	0.737	0.438	0.526	0.581	0.410

Table 4.2: Accuracy-tuned TabNet hyperparameters

Hyperparameter	**Value**	**Description**
width_prediction_and_attention	54	Width of decision prediction layer and attention embedding.
n_steps	10	Number of sequential decision steps.
gamma	1.6217	Feature re-use amount ($\gamma \geq 1$).
n_independent	4	Number of independent GLU blocks in the decoder.
n_shared	5	Number of shared GLU blocks in the decoder.
momentum	0.02599	Momentum for batch normalization.
lambda_sparse	0.0004755	Extra sparsity loss strength.
mask_type	entmax	Masking function type.
n_base_models	1	Number of base models for the ensemble.
n_epochs	8	Number of epochs determined by early stopping.

The predictive performance of the accuracy-tuned TabNet model is visualized as a

confusion matrix in Figure 4.1, where two out of eight actual positives and one out of eight actual negatives were misclassified.

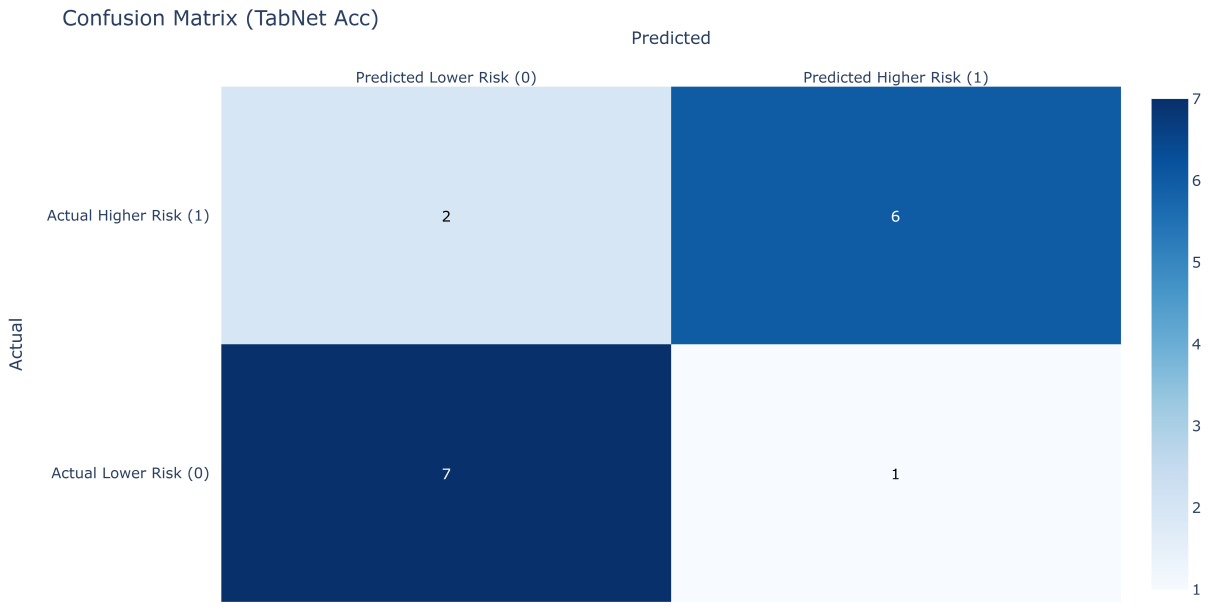


Figure 4.1: Confusion for accuracy-tuned TabNet

4.1.1.2 F1-Tuned TabNet

The best-performing F1-tuned models from each model family are included in Table 4.3. TabNet, despite relying solely the full feature set, achieves the highest validation performance across all model types, slightly beating the second candidate model by 0.014. It generalizes poorly, suffering a loss of 0.249 in F1-score when performing inference on the test set. This TabNet model, referred to as the *F1-tuned TabNet*, has the finalized hyperparameters as summarized and described in Table 4.4.

Table 4.3: Best-performing models with the highest validation F1-scores from each model family and their corresponding feature subsets. The integers 1 to 5 in the table represent different features - 1:CTS, 2: MTS, 3: LTS, 4:MTD, 5:sex

Model	**Features**	**Val F1**	**Test Acc**	**Test F1**	**Test F2**	**Test AUC**
TabNet	[1, 2, 3, 4, 5]	0.865	0.688	0.615	0.541	0.809
CatBoost	[1, 3, 4]	0.851	0.750	0.667	0.556	0.897
XGBoost	[1, 4, 5]	0.826	0.688	0.667	0.641	0.849
RF	[1, 3, 4]	0.770	0.750	0.667	0.556	0.825
MLP	[2, 3]	0.758	0.500	0.600	0.682	0.513
SVM	[3, 4]	0.735	0.812	0.824	0.854	0.791
KNN	[4, 5]	0.716	0.875	0.857	0.789	0.873

Table 4.4: F1-tuned TabNet hyperparameters

Hyperparameter	**Value**	**Description**
width_prediction_and_attention	45	Width of decision prediction layer and attention embedding.
n_steps	10	Number of sequential decision steps.
gamma	1.3066	Feature re-use amount ($\gamma \geq 1$).
n_independent	5	Number of independent GLU blocks in the decoder.
n_shared	4	Number of shared GLU blocks in the decoder.
momentum	0.01818	Momentum for batch normalization.
lambda_sparse	3.804×10^{-5}	Extra sparsity loss strength.
mask_type	sparsemax	Masking function type.
n_base_models	1	Number of base models for the ensemble.
n_epochs	12	Number of epochs determined by early stopping on validation subsets.

The predictive performance of the F1-tuned TabNet model is visualized as a confusion matrix in Figure 4.1, where four out of eight actual positives and one out of eight actual negatives were misclassified.

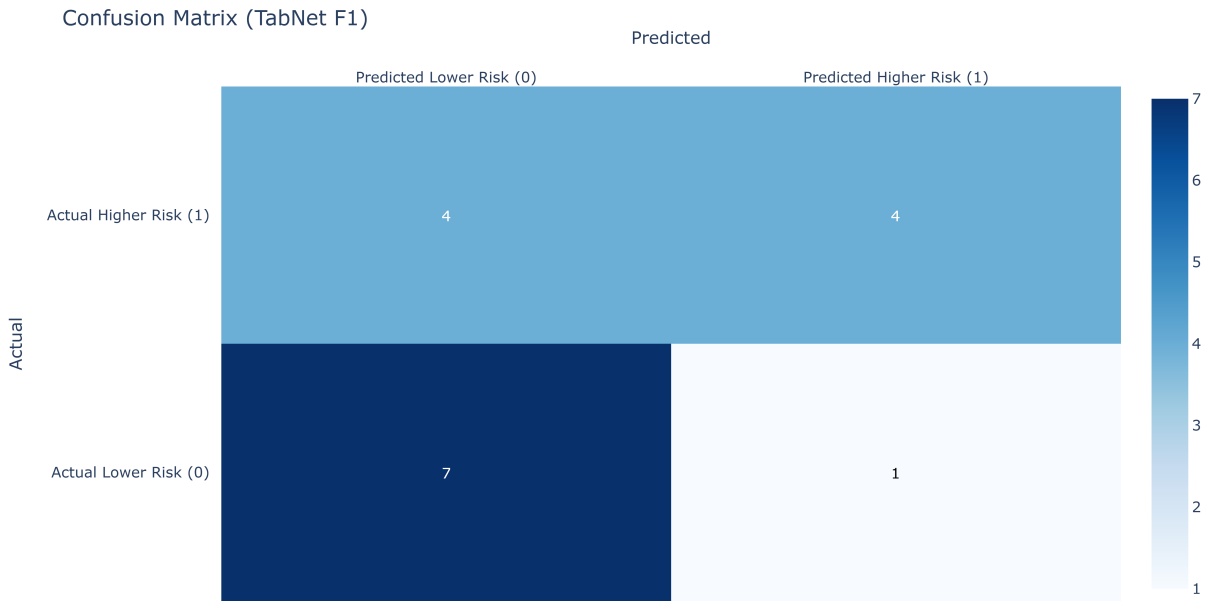


Figure 4.2: Confusion for F1-tuned TabNet

4.1.2 Global Feature Importance

The SHAP-enabled global feature importance of the best-performing models found from Section 4.1 are presented as SHAP beeswarm plots in Figures 4.3 and 4.5. These plots rank features by the mean absolute SHAP value while preserving the distribution of individual contributions. The vertical position, from top to bottom, represents the calculated order of feature importance. Each dot in the horizontal spread represents one patient, and its SHAP value quantifies the amount by which that features pushes the prediction away from the baseline. The color maps feature values to reveal interaction directionality.

Since TabNet also supports internal feature importance quantification through its masking mechanism, TabNet-calculated global feature importance o fthe best-performing models from Section 4.1 are presented in Figures 4.4 and 4.6. The feature importance score for each feature reflects the total attention allocated during sequential decision steps, computed as the sum of learned feature masks across all attentive transformer blocks and instances. Higher values indicate features that TabNet consistently selects and relies upon

for discrimination whereas lower values indicate features less attended to by TabNet.

While two sets of results are presented, this thesis focuses on SHAP FIA for fair comparison throughout the study.

4.1.2.1 Accuracy-Tuned TabNet

Figure 4.3 shows that the global SHAP values computed from the accuracy-tuned TabNet, in descending order, are 0.2184 (sex), 0.2043 (LTS), 0.1995 (MTD), 0.1887 (CTS), and 0.1720 (MTS). Despite the higher predictive performance demonstrated in Figure 4.1, no distinguishable clusters were identified in the beeswarm plot, as data points with high and low feature values all remain unseparated.

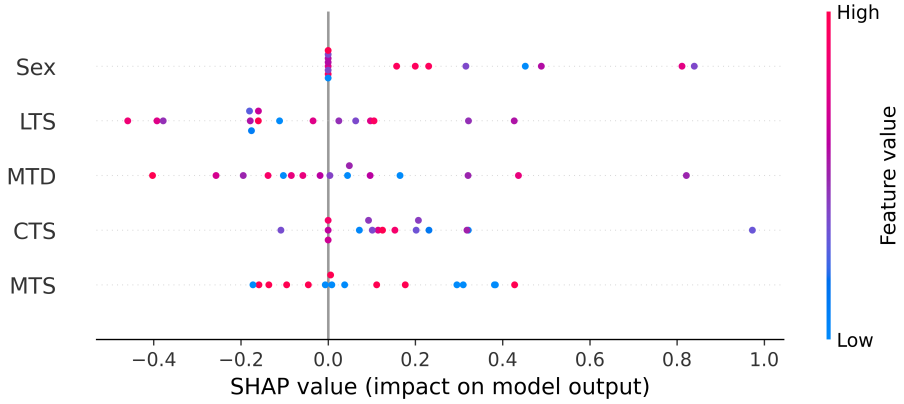


Figure 4.3: Beeswarm plot for accuracy-tuned TabNet

Figure 4.4 shows that the global TabNet feature importance scores computed from the accuracy-tuned TabNet, in descending order, are 0.2874 (LTS), 0.2223 (MTD), 0.1984 (CTS), 0.1711 (MTS), and 0.1209 (sex). The feature importance scores computed here is normalized, and therefore add up to a sum greater than the sum of global SHAP values. Similar operations can be applied to SHAP values, but is deemed unnecessary as the main interest lies in the ranking order and the relative relationships.

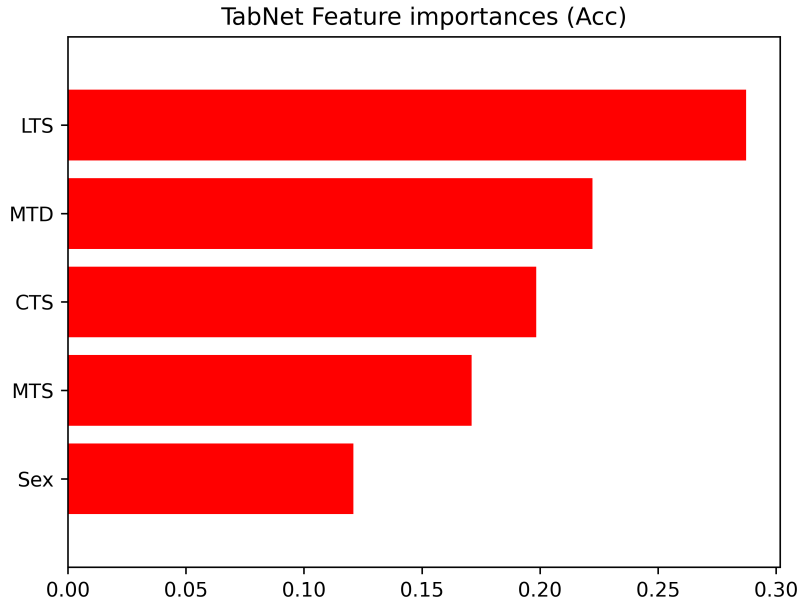


Figure 4.4: Internal feature importance of accuracy-tuned TabNet

4.1.2.2 F1-Tuned TabNet

Figure 4.5 shows that the global SHAP values computed from the F1-tuned TabNet, in descending order, are 0.2282 (LTS), 0.2157 (MTS), 0.1337 (CTS), 0.0679 (MTD), and 0.0657 (sex). Despite the higher predictive performance demonstrated in Figure 4.1, Similar to the accuracy-tuned counterpart, no distinguishable clusters were identified in the beeswarm plot, as data points with high and low feature values all remain unseparated.

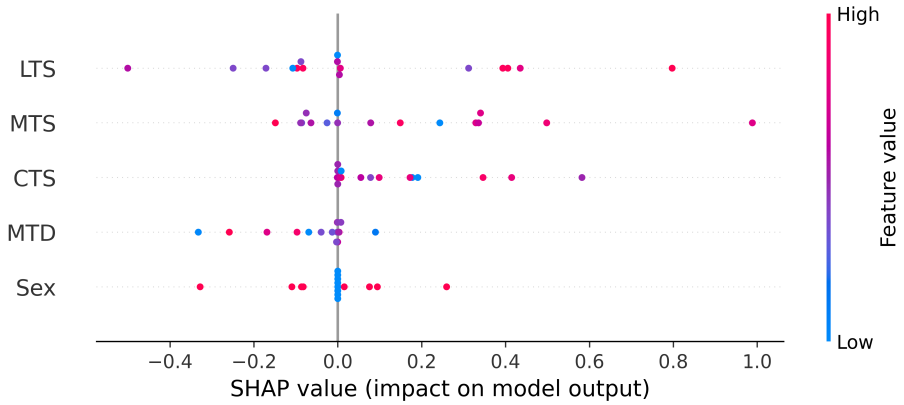


Figure 4.5: Beeswarm plot for F1-tuned TabNet

Figure 4.4 shows that the global TabNet feature importance scores computed from the accuracy-tuned TabNet, in descending order, are 0.2612 (MTS), 0.2572 (MTD), 0.1996 (CTS), 0.1947 (LTS), and 0.0869 (sex).

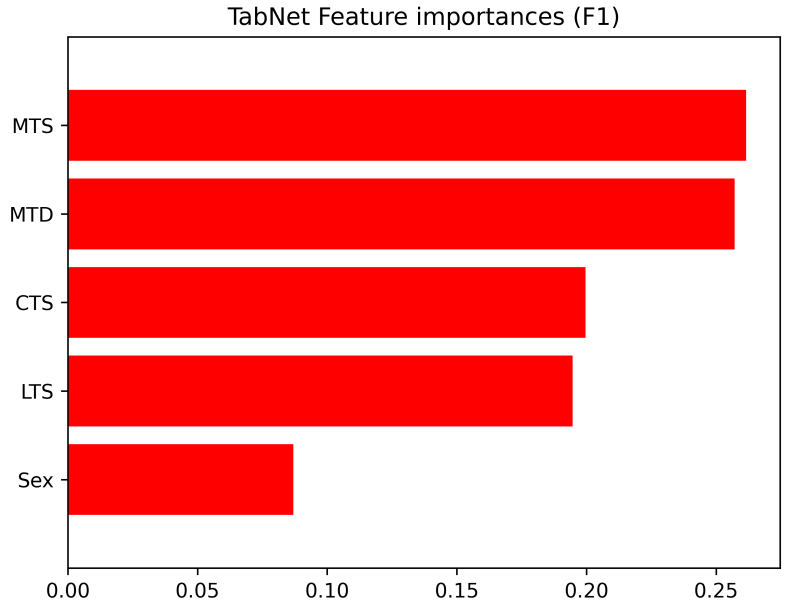


Figure 4.6: Internal feature importance of F1-tuned TabNet

4.2 AutoGluon Pipeline

Since models in the AutoGluon pipeline are cheaper to train and experiment with, they are tuned using all three intended metrics, namely balanced accuracy, the F1 score, and the F2 score. It should be noted that $WeightedEnsemble\_L2$ represent the super-model obtained through multi-layer stack ensemble, consisting of a subset of the Layer 1 (L1) bag models. When discussing on the performance level, $WeightedEnsemble\_L2$ is referred to as a model, but mechanistically it is the weighted sum of multiple L1 bag models.

4.2.1 Test performance

The best-average-performing model from each of the model families are recorded and tabulated in Tables 4.5, 4.7, and 4.9.

For transparency, the details of the best-average-performing models from all seeds are shown in Tables 4.6 to 4.10. This is particularly important, as the standard deviations observed in Tables 4.5 to 4.9 are not negligible. A specific instance taken from the ten seeds has the chance of under-performing or over-performing the reported average scores.

The confusion matrices corresponding to the best-average-performing models are presented in Figure 4.7, 4.8, and 4.9 for visualization of predictive performance and transparency, as it can be used to derive other unreported metrics.

4.2.1.1 Balanced-Accuracy-Tuned Standard Ensemble

The best-performing accuracy-tuned models from each model family are included in Table 4.5. The best-performing model across all model families and variants is the $WeightedEnsemble\_L2$, slightly beating the second candidate model by 0.012. It generalizes decently well, suffering an average loss of 0.100 in balanced accuracy when performing inference on the test set. The standard deviation of test balanced accuracy is reported to be 0.117, which was computed from Table 4.6. The instance of $WeightedEnsemble\_L2$ from the default seed of 42, referred to as the balanced-accuracy-tuned standard ensemble, consists of three different L1 bag models: $CatBoost\_r24\_BAG\_L1$ (weight = 0.875), $TabM\_r184\_BAG\_L1$ (weight = 0.0625), and $XGBoost\_r40\_BAG\_L1$ (weight = 0.0625).

Table 4.5: Performance of balanced-accuracy-tuned models ranked by validation score. Metrics are averaged across ten runs

Model	Val Bal. Acc	Test Bal. Acc	Test F1	Test F2	Test AUC	Test Precision	Test Recall
WeightedEnsemble_L2	0.811 ± 0.032	0.711 ± 0.117	0.675 ± 0.141	0.652 ± 0.158	0.795 ± 0.084	0.744 ± 0.159	0.640 ± 0.171
TabM_BAG_L1	0.799 ± 0.035	0.712 ± 0.114	0.682 ± 0.140	0.673 ± 0.169	0.795 ± 0.102	0.714 ± 0.110	0.670 ± 0.189
TabM_r49_BAG_L1	0.793 ± 0.043	0.698 ± 0.124	0.670 ± 0.148	0.662 ± 0.168	0.805 ± 0.098	0.698 ± 0.135	0.660 ± 0.184
CatBoost_r91_BAG_L1	0.778 ± 0.029	0.708 ± 0.120	0.684 ± 0.136	0.674 ± 0.147	0.789 ± 0.122	0.708 ± 0.132	0.670 ± 0.157
CatBoost_r24_BAG_L1	0.764 ± 0.031	0.712 ± 0.131	0.681 ± 0.151	0.666 ± 0.174	0.785 ± 0.116	0.737 ± 0.165	0.660 ± 0.190
CatBoost_r10_BAG_L1	0.762 ± 0.028	0.697 ± 0.112	0.669 ± 0.121	0.650 ± 0.132	0.765 ± 0.134	0.716 ± 0.149	0.640 ± 0.143
CatBoost_BAG_L1	0.762 ± 0.034	0.721 ± 0.121	0.690 ± 0.144	0.670 ± 0.158	0.763 ± 0.142	0.746 ± 0.170	0.660 ± 0.171
CatBoost_r51_BAG_L1	0.759 ± 0.038	0.710 ± 0.134	0.668 ± 0.164	0.637 ± 0.171	0.776 ± 0.129	0.741 ± 0.157	0.620 ± 0.175
TabM_r52_BAG_L1	0.752 ± 0.030	0.690 ± 0.089	0.632 ± 0.130	0.591 ± 0.150	0.777 ± 0.110	0.745 ± 0.148	0.570 ± 0.164
LightGBM_r11_BAG_L1	0.749 ± 0.031	0.663 ± 0.107	0.559 ± 0.198	0.507 ± 0.207	0.787 ± 0.132	0.750 ± 0.179	0.480 ± 0.210
TabM_r69_BAG_L1	0.747 ± 0.028	0.670 ± 0.090	0.603 ± 0.130	0.561 ± 0.157	0.789 ± 0.110	0.744 ± 0.157	0.540 ± 0.171
TabM_r184_BAG_L1	0.744 ± 0.036	0.670 ± 0.127	0.600 ± 0.199	0.567 ± 0.210	0.754 ± 0.121	0.708 ± 0.213	0.550 ± 0.217
XGBoost_r40_BAG_L1	0.741 ± 0.031	0.696 ± 0.104	0.607 ± 0.166	0.543 ± 0.183	0.795 ± 0.121	0.785 ± 0.122	0.510 ± 0.191
LightGBM_r33_BAG_L1	0.734 ± 0.053	0.686 ± 0.087	0.598 ± 0.137	0.533 ± 0.158	0.798 ± 0.118	0.796 ± 0.144	0.500 ± 0.170
TabM_r191_BAG_L1	0.732 ± 0.033	0.664 ± 0.093	0.581 ± 0.151	0.534 ± 0.174	0.757 ± 0.098	0.738 ± 0.144	0.510 ± 0.185
XGBoost_r171_BAG_L1	0.722 ± 0.041	0.691 ± 0.099	0.599 ± 0.165	0.534 ± 0.178	0.793 ± 0.112	0.783 ± 0.143	0.500 ± 0.183
Mitra_BAG_L1	0.710 ± 0.042	0.725 ± 0.104	0.693 ± 0.124	0.671 ± 0.143	0.792 ± 0.111	0.752 ± 0.126	0.660 ± 0.158
TabICL_BAG_L1	0.678 ± 0.051	0.763 ± 0.087	0.733 ± 0.112	0.706 ± 0.127	0.778 ± 0.100	0.793 ± 0.097	0.690 ± 0.137
TabPFNv2_r143_BAG_L1	0.672 ± 0.030	0.672 ± 0.109	0.624 ± 0.143	0.601 ± 0.170	0.774 ± 0.111	0.690 ± 0.112	0.590 ± 0.185
TabPFNv2_r94_BAG_L1	0.665 ± 0.041	0.668 ± 0.117	0.617 ± 0.149	0.598 ± 0.188	0.791 ± 0.115	0.701 ± 0.154	0.590 ± 0.213
TabPFNv2_r181_BAG_L1	0.662 ± 0.036	0.667 ± 0.102	0.621 ± 0.127	0.594 ± 0.144	0.780 ± 0.095	0.695 ± 0.120	0.580 ± 0.155
LightGBM_r21_BAG_L1	0.500 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 4.6 reveals the performance of all instances of balanced-accuracy-tuned ensembles obtained from the ten seeds. According to the tabulated values, the model obtained from the default seed of 42 is favored by an easier data split, resulting in highest test performances across the entire table. It achieves a test F1-score of 0.947 which is 0.173 higher than the validation F-1 score.

Table 4.6: Performance metrics for F2-tuned Ensemble different random seeds

Random Seed	Val Bal. Acc	Test Bal. Acc	Test F1	Test F2	Test ROC AUC	Test Precision	Test Recall
42	0.777	0.950	0.947	0.918	0.973	1.000	0.900
52	0.786	0.750	0.667	0.556	0.827	1.000	0.500
62	0.782	0.855	0.842	0.816	0.818	0.889	0.800
72	0.828	0.659	0.588	0.532	0.736	0.714	0.500
82	0.799	0.668	0.667	0.686	0.773	0.636	0.700
92	0.804	0.718	0.727	0.769	0.809	0.667	0.800
102	0.812	0.664	0.632	0.612	0.736	0.667	0.600
112	0.879	0.573	0.571	0.588	0.636	0.545	0.600
122	0.812	0.609	0.500	0.435	0.782	0.667	0.400
132	0.838	0.714	0.700	0.700	0.827	0.700	0.700

The predictive performance of the balanced-accuracy-tuned AutoGluon model is visualized as a confusion matrix in Figure 4.7, where one out of ten actual positives and none out of eleven actual negatives were misclassified.

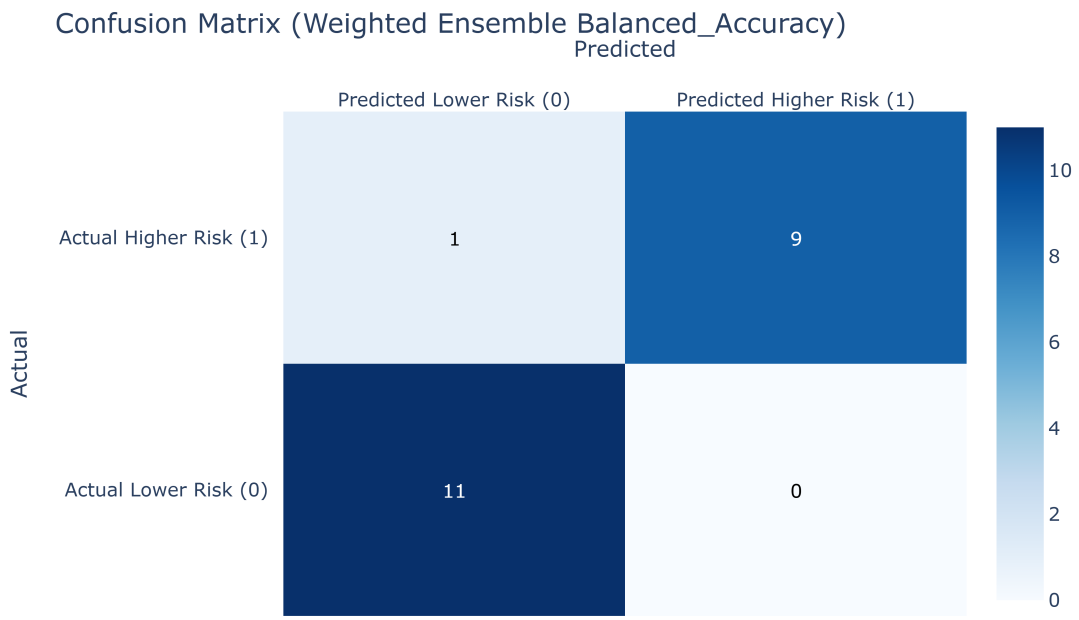


Figure 4.7: Confusion matrix for the balanced-accuracy-tuned ensemble

4.2.1.2 F1-Tuned Standard Ensemble

The best-average-performing F1-tuned models from each model family are included in Table 4.7. The best-average-performing model across all model families and variants is the $WeightedEnsemble\_L2$, slightly beating the second candidate model by 0.010. It generalizes decently well, suffering an average loss of 0.092 in F1 score when performing inference on the test set. The standard deviation of test F1-score is reported to be 0.110, which was computed from Table 4.8. The instance of $WeightedEnsemble\_L2$ from the default seed of 42, referred to as the F1-tuned standard ensemble, consists of two different L1 models: $TabM\_r49\_BAG\_L1$ (weight = 0.75) and $LightGBM\_r11\_BAG\_L1$ (weight = 0.25).

Table 4.7: Performance of F1-tuned models ranked by validation score. Metrics are averaged across ten runs

Model	Val F1	Test Bal. Acc	Test F1	Test F2	Test AUC	Test Precision	Test Recall
WeightedEnsemble_L2	0.788 ± 0.044	0.710 ± 0.100	0.696 ± 0.110	0.703 ± 0.130	0.795 ± 0.095	0.690 ± 0.092	0.710 ± 0.145
TabM_BAG_L1	0.778 ± 0.043	0.709 ± 0.105	0.688 ± 0.122	0.688 ± 0.148	0.798 ± 0.096	0.701 ± 0.107	0.690 ± 0.166
TabM_r49_BAG_L1	0.775 ± 0.045	0.700 ± 0.112	0.680 ± 0.132	0.685 ± 0.156	0.793 ± 0.098	0.680 ± 0.102	0.690 ± 0.173
TabM_r69_BAG_L1	0.711 ± 0.046	0.676 ± 0.102	0.627 ± 0.127	0.595 ± 0.150	0.785 ± 0.109	0.733 ± 0.168	0.580 ± 0.169
TabM_r52_BAG_L1	0.709 ± 0.045	0.668 ± 0.109	0.627 ± 0.134	0.602 ± 0.143	0.779 ± 0.110	0.691 ± 0.149	0.590 ± 0.152
TabM_r184_BAG_L1	0.707 ± 0.048	0.690 ± 0.128	0.633 ± 0.199	0.605 ± 0.206	0.760 ± 0.121	0.712 ± 0.216	0.590 ± 0.213
LightGBM_r11_BAG_L1	0.702 ± 0.047	0.668 ± 0.105	0.572 ± 0.185	0.518 ± 0.193	0.788 ± 0.134	0.750 ± 0.179	0.490 ± 0.197
LightGBM_r33_BAG_L1	0.694 ± 0.074	0.707 ± 0.103	0.642 ± 0.149	0.589 ± 0.163	0.787 ± 0.108	0.782 ± 0.152	0.560 ± 0.171
XGBoost_r40_BAG_L1	0.693 ± 0.047	0.707 ± 0.096	0.634 ± 0.154	0.579 ± 0.173	0.795 ± 0.121	0.781 ± 0.123	0.550 ± 0.184
TabM_r191_BAG_L1	0.683 ± 0.045	0.690 ± 0.092	0.631 ± 0.128	0.591 ± 0.151	0.763 ± 0.104	0.739 ± 0.108	0.570 ± 0.164
Mitra_BAG_L1	0.672 ± 0.053	0.730 ± 0.116	0.695 ± 0.138	0.672 ± 0.161	0.799 ± 0.103	0.758 ± 0.132	0.660 ± 0.178
XGBoost_r171_BAG_L1	0.659 ± 0.064	0.706 ± 0.107	0.625 ± 0.178	0.569 ± 0.197	0.795 ± 0.113	0.783 ± 0.145	0.540 ± 0.207
CatBoost_r91_BAG_L1	0.657 ± 0.046	0.697 ± 0.096	0.663 ± 0.118	0.642 ± 0.126	0.783 ± 0.129	0.711 ± 0.120	0.630 ± 0.134
CatBoost_BAG_L1	0.644 ± 0.049	0.721 ± 0.113	0.690 ± 0.133	0.670 ± 0.151	0.774 ± 0.124	0.744 ± 0.148	0.660 ± 0.165
CatBoost_r51_BAG_L1	0.641 ± 0.050	0.706 ± 0.112	0.671 ± 0.138	0.650 ± 0.156	0.759 ± 0.137	0.731 ± 0.144	0.640 ± 0.171
CatBoost_r10_BAG_L1	0.637 ± 0.050	0.697 ± 0.102	0.660 ± 0.129	0.640 ± 0.148	0.766 ± 0.132	0.716 ± 0.130	0.630 ± 0.164
TabICL_BAG_L1	0.633 ± 0.066	0.763 ± 0.087	0.733 ± 0.112	0.706 ± 0.127	0.778 ± 0.100	0.793 ± 0.097	0.690 ± 0.137
CatBoost_r24_BAG_L1	0.627 ± 0.050	0.698 ± 0.122	0.670 ± 0.147	0.662 ± 0.168	0.755 ± 0.138	0.697 ± 0.126	0.660 ± 0.184
TabPFNv2_r143_BAG_L1	0.625 ± 0.051	0.672 ± 0.109	0.624 ± 0.143	0.601 ± 0.170	0.774 ± 0.111	0.690 ± 0.112	0.590 ± 0.185
TabPFNv2_r94_BAG_L1	0.620 ± 0.052	0.668 ± 0.117	0.617 ± 0.149	0.598 ± 0.188	0.791 ± 0.115	0.701 ± 0.154	0.590 ± 0.213
TabPFNv2_r181_BAG_L1	0.616 ± 0.050	0.667 ± 0.102	0.621 ± 0.127	0.594 ± 0.144	0.780 ± 0.095	0.695 ± 0.120	0.580 ± 0.155
LightGBM_r21_BAG_L1	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 4.8 reveals the performance of all instances of F1-tuned AutoGluon ensembles obtained from the ten seeds. According to the tabulated values, the model obtained from the default seed of 42 is favored by an easier data split, resulting in highest test performances across the entire table. It achieves a test F1-score of 0.952 which is 0.227 higher than the validation F-1 score.

Table 4.8: Performance metrics for F1-tuned Ensemble different random seeds

Random Seed	Val F1	Test Bal. Acc	Test F1	Test F2	Test ROC AUC	Test Precision	Test Recall
42	0.725	0.955	0.952	0.980	1.000	0.909	1.000
52	0.725	0.714	0.700	0.700	0.818	0.700	0.700
62	0.773	0.718	0.727	0.769	0.845	0.667	0.800
72	0.805	0.659	0.588	0.532	0.736	0.714	0.500
82	0.771	0.668	0.667	0.686	0.782	0.636	0.700
92	0.785	0.718	0.727	0.769	0.764	0.667	0.800
102	0.789	0.664	0.632	0.612	0.727	0.667	0.600
112	0.872	0.573	0.571	0.588	0.636	0.545	0.600
122	0.811	0.664	0.632	0.612	0.800	0.667	0.600
132	0.827	0.764	0.762	0.784	0.845	0.727	0.800

The predictive performance of the F1-tuned AutoGluon model is visualized as a confusion matrix in Figure 4.8, where none out of ten actual positives and one out of eleven actual negatives were misclassified.

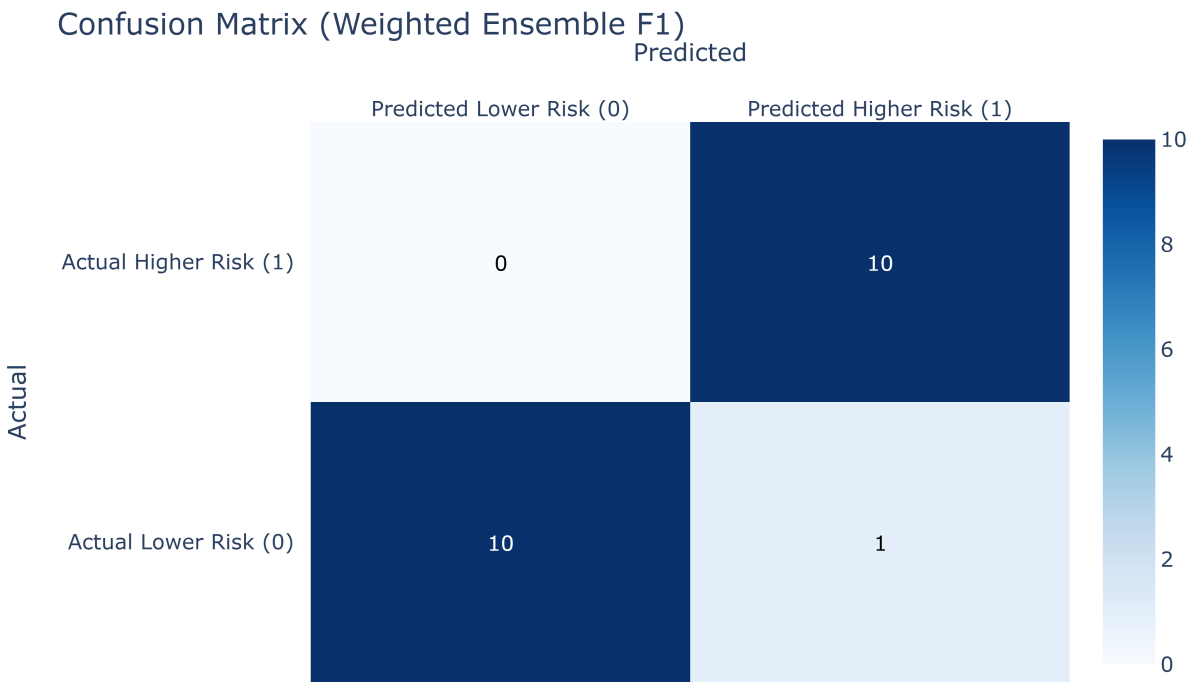


Figure 4.8: Confusion matrix for the F1-tuned ensemble

4.2.1.3 F2-Tuned Standard Ensemble

The best-average-performing F2-tuned models from each model family are included in Table 4.9. The best-average-performing model across all model families and variants is the $WeightedEnsemble\_L2$, slightly beating the second candidate model by 0.007. It generalizes well, suffering an average loss of 0.048 in F2 score when performing inference on the test set. The standard deviation of test F2-score is reported to be 0.128, which was computed from Table 4.10. The instance of $WeightedEnsemble\_L2$ from the default seed of 42, referred to as the F2-tuned standard ensemble, consists of two different L1 models: $TabM\_r49\_BAG\_L1$ (weight = 0.75) and $LightGBM\_r11\_BAG\_L1$ (weight = 0.25).

Table 4.9: Performance of F2-tuned ranked by validation score. Metrics are averaged across ten runs

Model	Val F2	Test Bal. Acc	Test F1	Test F2	Test AUC	Test Precision	Test Recall
WeightedEnsemble_L2	0.786 ± 0.052	0.734 ± 0.091	0.724 ± 0.100	0.738 ± 0.128	0.807 ± 0.096	0.716 ± 0.099	0.750 ± 0.151
TabM_r49_BAG_L1	0.779 ± 0.052	0.725 ± 0.090	0.714 ± 0.101	0.728 ± 0.128	0.801 ± 0.095	0.703 ± 0.086	0.740 ± 0.151
TabM_BAG_L1	0.764 ± 0.047	0.704 ± 0.108	0.684 ± 0.124	0.686 ± 0.148	0.803 ± 0.097	0.694 ± 0.111	0.690 ± 0.166
TabM_r69_BAG_L1	0.686 ± 0.067	0.686 ± 0.100	0.642 ± 0.127	0.613 ± 0.151	0.783 ± 0.110	0.740 ± 0.165	0.600 ± 0.170
TabM_r184_BAG_L1	0.680 ± 0.063	0.676 ± 0.124	0.609 ± 0.204	0.583 ± 0.221	0.762 ± 0.122	0.704 ± 0.215	0.570 ± 0.231
TabM_r52_BAG_L1	0.669 ± 0.066	0.663 ± 0.107	0.620 ± 0.131	0.594 ± 0.139	0.784 ± 0.108	0.688 ± 0.150	0.580 ± 0.148
LightGBM_r33_BAG_L1	0.659 ± 0.085	0.707 ± 0.103	0.642 ± 0.149	0.589 ± 0.163	0.788 ± 0.109	0.782 ± 0.152	0.560 ± 0.171
XGBoost_r40_BAG_L1	0.647 ± 0.059	0.702 ± 0.094	0.630 ± 0.151	0.577 ± 0.171	0.795 ± 0.121	0.774 ± 0.127	0.550 ± 0.184
LightGBM_r11_BAG_L1	0.645 ± 0.057	0.668 ± 0.105	0.572 ± 0.185	0.518 ± 0.193	0.790 ± 0.133	0.750 ± 0.179	0.490 ± 0.197
CatBoost_r91_BAG_L1	0.639 ± 0.066	0.697 ± 0.096	0.663 ± 0.118	0.642 ± 0.126	0.783 ± 0.129	0.711 ± 0.120	0.630 ± 0.134
Mitra_BAG_L1	0.637 ± 0.075	0.730 ± 0.113	0.695 ± 0.139	0.672 ± 0.162	0.795 ± 0.110	0.750 ± 0.124	0.660 ± 0.178
TabM_r191_BAG_L1	0.635 ± 0.058	0.690 ± 0.092	0.631 ± 0.128	0.591 ± 0.151	0.765 ± 0.102	0.739 ± 0.108	0.570 ± 0.164
CatBoost_BAG_L1	0.625 ± 0.057	0.721 ± 0.113	0.690 ± 0.133	0.670 ± 0.151	0.774 ± 0.124	0.744 ± 0.148	0.660 ± 0.165
CatBoost_r10_BAG_L1	0.622 ± 0.062	0.697 ± 0.102	0.660 ± 0.129	0.640 ± 0.148	0.766 ± 0.132	0.716 ± 0.130	0.630 ± 0.164
CatBoost_r51_BAG_L1	0.620 ± 0.063	0.706 ± 0.112	0.671 ± 0.138	0.650 ± 0.156	0.759 ± 0.137	0.731 ± 0.144	0.640 ± 0.171
CatBoost_r24_BAG_L1	0.614 ± 0.059	0.698 ± 0.122	0.670 ± 0.147	0.662 ± 0.168	0.755 ± 0.138	0.697 ± 0.126	0.660 ± 0.184
TabICL_BAG_L1	0.603 ± 0.072	0.763 ± 0.087	0.733 ± 0.112	0.706 ± 0.127	0.778 ± 0.100	0.793 ± 0.097	0.690 ± 0.137
XGBoost_r171_BAG_L1	0.599 ± 0.085	0.706 ± 0.107	0.625 ± 0.178	0.569 ± 0.197	0.794 ± 0.113	0.783 ± 0.145	0.540 ± 0.207
TabPFNv2_r143_BAG_L1	0.596 ± 0.069	0.672 ± 0.109	0.624 ± 0.143	0.601 ± 0.170	0.774 ± 0.111	0.690 ± 0.112	0.590 ± 0.185
TabPFNv2_r94_BAG_L1	0.593 ± 0.060	0.668 ± 0.117	0.617 ± 0.149	0.598 ± 0.188	0.791 ± 0.115	0.701 ± 0.154	0.590 ± 0.213
TabPFNv2_r181_BAG_L1	0.588 ± 0.063	0.667 ± 0.102	0.621 ± 0.127	0.594 ± 0.144	0.780 ± 0.095	0.695 ± 0.120	0.580 ± 0.155
LightGBM_r21_BAG_L1	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 4.10 reveals the performance of all instances of F2-tuned AutoGluon ensembles obtained from the ten seeds. According to the tabulated values, the model obtained from the default seed of 42 was favored by an easier data split, resulting in the highest test performances across the entire table. It achieved a test F2-score of 0.980 which is 0.244 higher than the validation F-2 score.

Table 4.10: Performance metrics for different random seeds (F2-focused)

Random Seed	Val F2	Test Bal. Acc	Test F1	Test F2	Test ROC AUC	Test Precision	Test Recall
42	0.736	0.955	0.952	0.980	1.000	0.909	1.000
52	0.688	0.755	0.706	0.638	0.873	0.857	0.600
62	0.775	0.718	0.727	0.769	0.845	0.667	0.800
72	0.804	0.659	0.588	0.532	0.718	0.714	0.500
82	0.771	0.764	0.762	0.784	0.827	0.727	0.800
92	0.804	0.718	0.727	0.769	0.745	0.667	0.800
102	0.780	0.723	0.750	0.833	0.745	0.643	0.900
112	0.888	0.623	0.636	0.673	0.655	0.583	0.700
122	0.812	0.664	0.632	0.612	0.818	0.667	0.600
132	0.803	0.764	0.762	0.784	0.845	0.727	0.800

The predictive performance of the F2-tuned AutoGluon model is visualized as a confusion matrix in Figure 4.9, where none out of ten actual positives and one out of eleven actual negatives were misclassified.

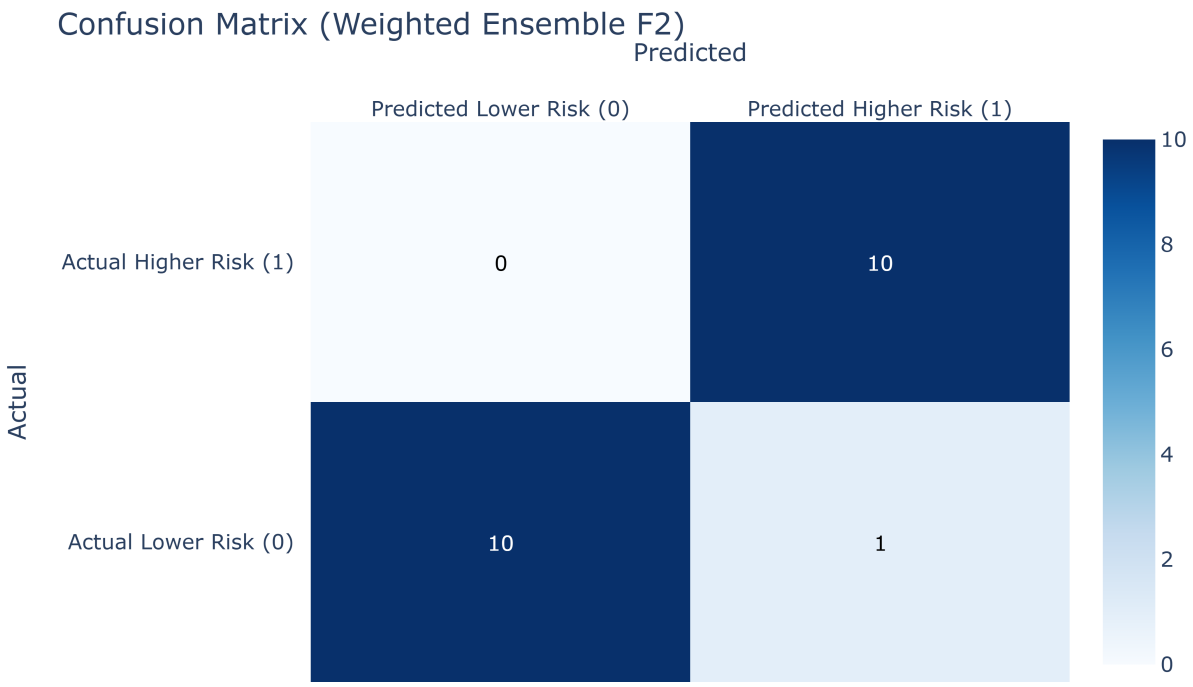


Figure 4.9: Confusion matrix for the F2-tuned ensemble

4.2.2 Global Feature Importance

Global SHAP values are measured using the same method as in Section 4.1.2. The beeswarm plots for all ten seeds are combined and presented in Figures 4.11, 4.15, and 4.15. Given that each seed could rank the feature importance differently and that the presented models are from the default seed of 42, the average SHAP value magnitudes are computed and compared against the SHAP value magnitudes obtained from seed 42. The comparisons can be found in Figures 4.10, 4.12, and 4.14.

4.2.2.1 Balanced-Accuracy-Tuned Standard Ensemble

Figure 4.10 shows that the global SHAP values computed from the balanced-accuracy-tuned standard ensemble are consistent with the average global SHAP values computed from the balanced-accuracy-tuned ensembles obtained across all ten seeds. The former, in descending order, are: 0.0280 (MTD), 0.0181 (LTS), 0.0116 (MTS), 0.0068 (sex), and 0.0066 (CTS). The latter, in descending order, are: 0.1093 (MTD), 0.0399 (LTS), 0.0267 (MTS), 0.0168 (sex), and 0.0168 (CTS).

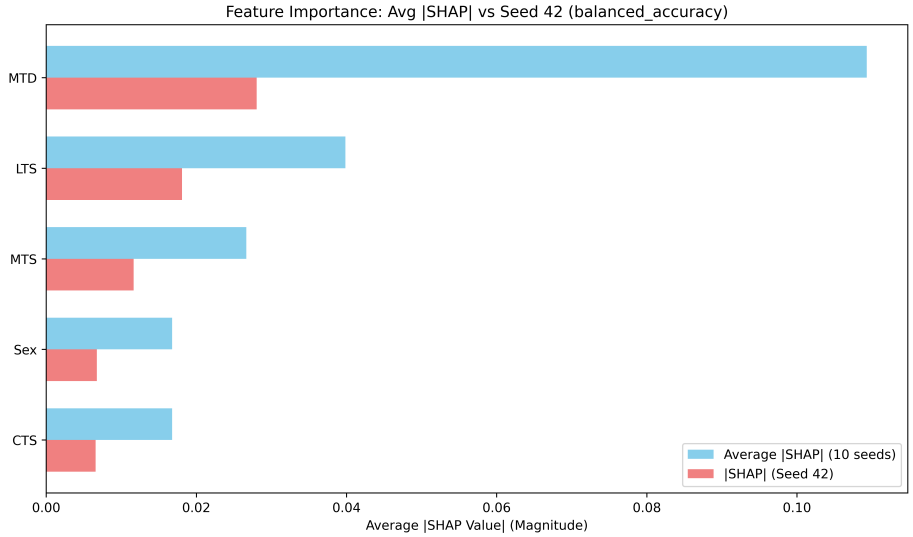


Figure 4.10: Averaged global feature importance by balanced-accuracy-tuned ensemble

Figure 4.11 provides clear separation of the clusters for features *MTD*, *LTS*, *MTS*, and *sex* based on their relative magnitudes. Regardless of the ordering of feature importance, decreased MTD, increased LTS and MTS, and being male are all associated with higher SHAP values contributing to predicted positives.

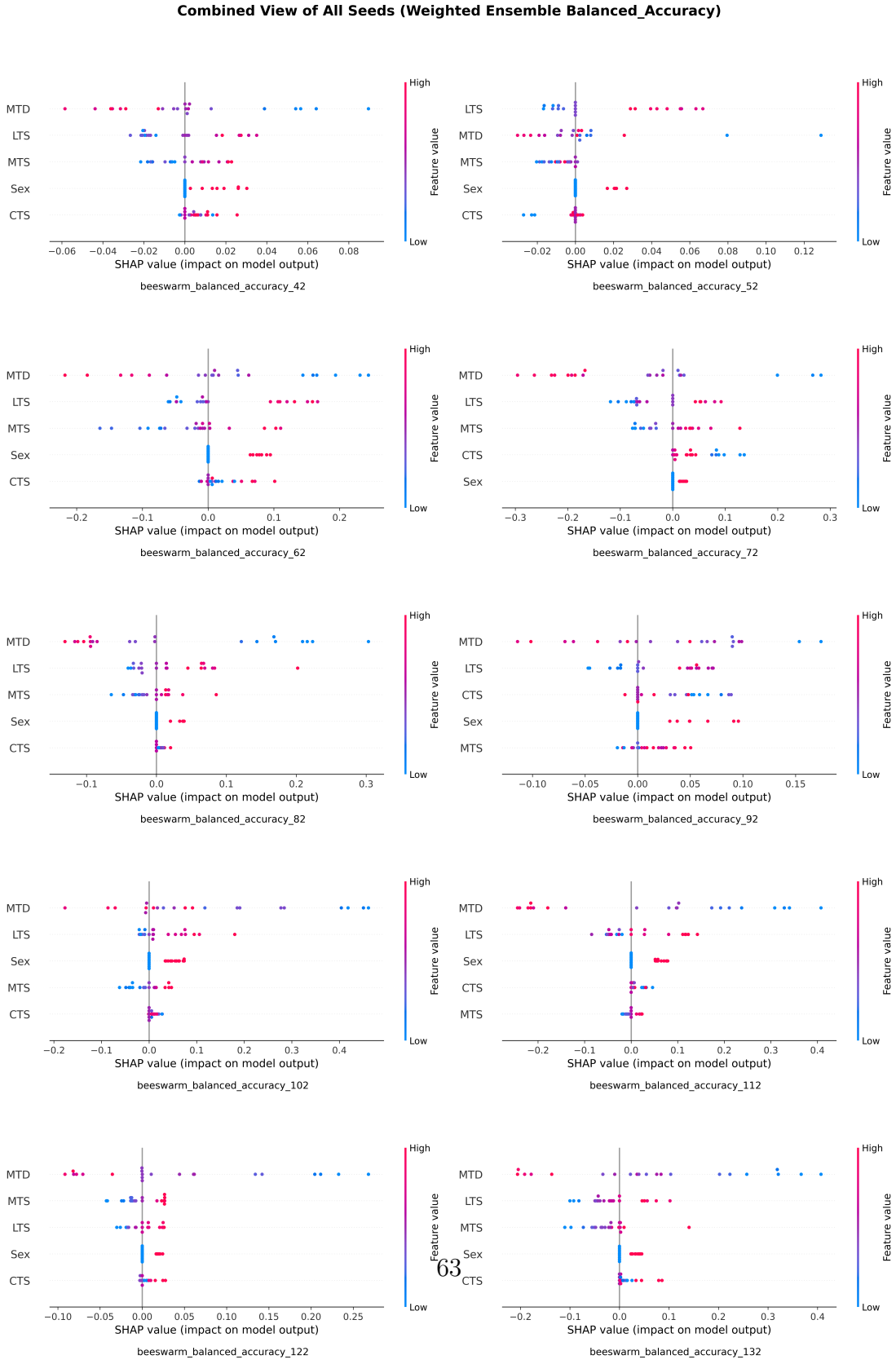


Figure 4.11: Combined view of global feature importance by balanced-accuracy-tuned ensemble

4.2.2.2 F1-Tuned Standard Ensemble

Figure 4.12 shows that the global SHAP values computed from the balanced-accuracy-tuned standard ensemble are mostly consistent with the average global SHAP values computed from the balanced-accuracy-tuned ensembles obtained across all ten seeds, except for the ordering of CTS and sex. The former, in descending order, are: 0.1216 (MTD), 0.0383 (LTS), 0.0369 (MTS), 0.0210 (CTS), and 0.0129 (sex). The latter, in descending order, are: 0.1544 (MTD), 0.0425 (LTS), 0.0320 (MTS), 0.0205 (sex), and 0.0178 (CTS).

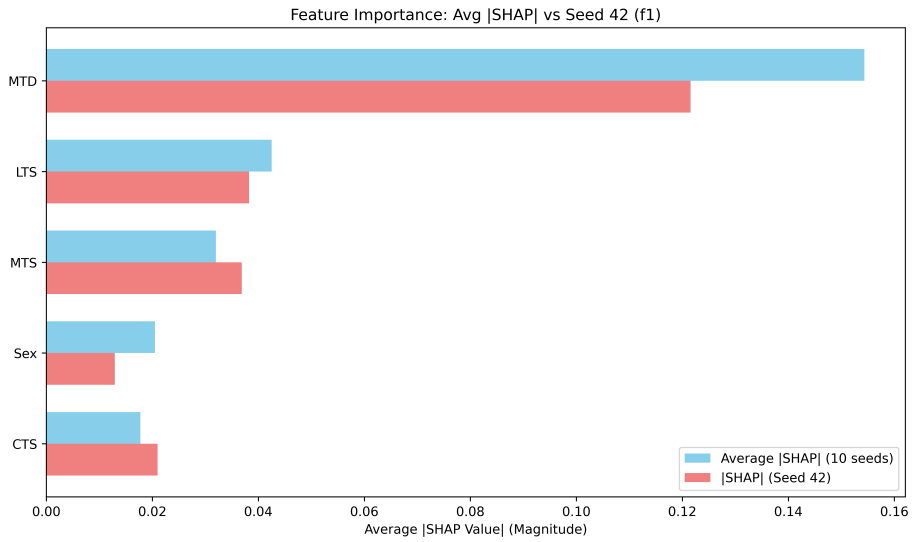


Figure 4.12: Averaged global feature importance by F1-tuned ensemble

Figure 4.13 provides clear separation of the clusters for features *MTD*, *LTS*, *MTS*, and *sex* based on their relative magnitudes. Regardless of the ordering of feature importance, decreased MTD, increased LTS and MTS, and being male are all associated with higher SHAP values contributing to predicted positives.

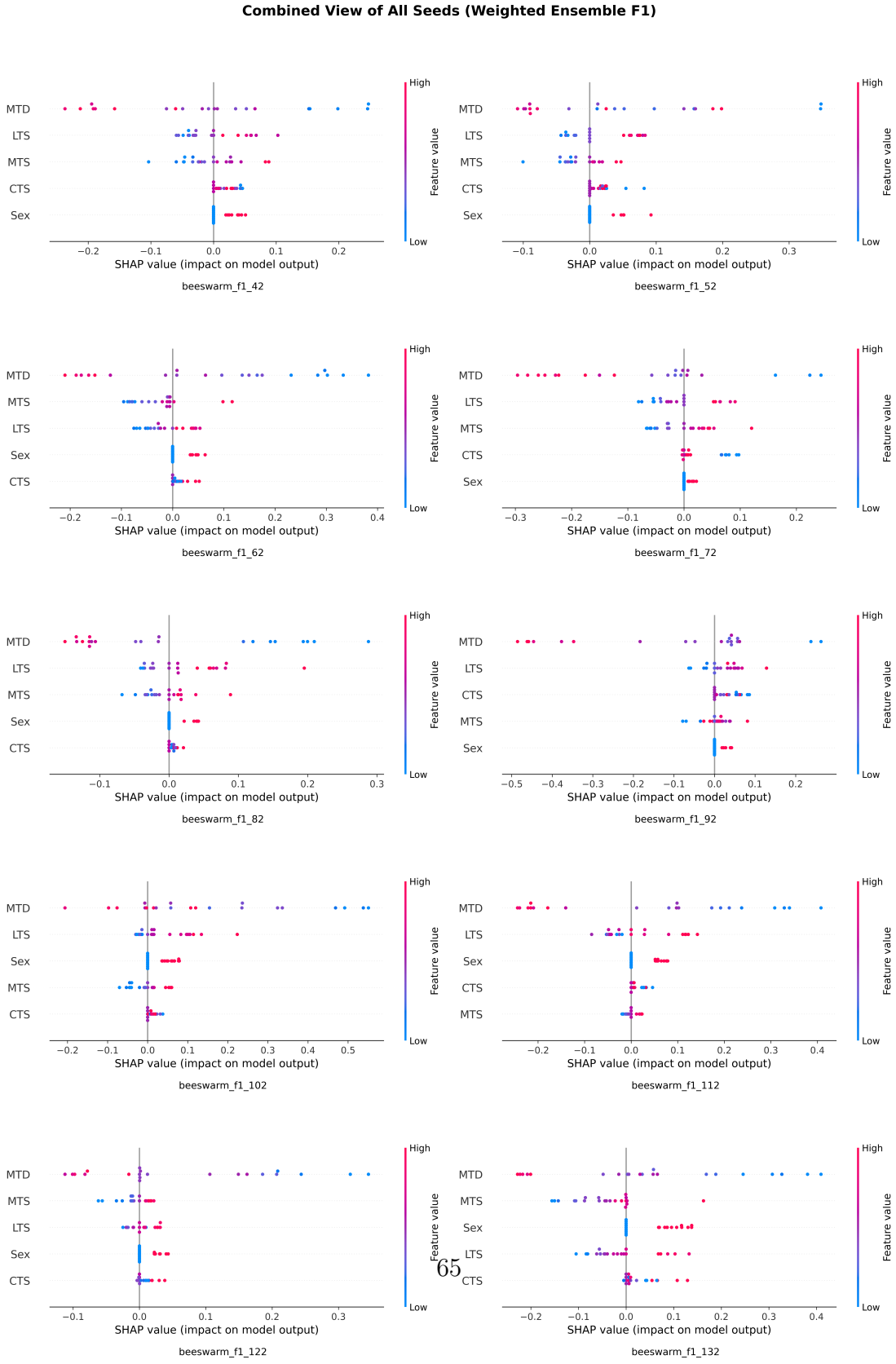


Figure 4.13: Combined view of global feature importance by F1-tuned ensemble

4.2.2.3 F2-Tuned Standard Ensemble

Figure 4.14 shows that the global SHAP values computed from the balanced-accuracy-tuned standard ensemble are consistent with the average global SHAP values computed from the balanced-accuracy-tuned ensembles obtained across all ten seeds. The former, in descending order, are: 0.1216 (MTD), 0.0383 (LTS), 0.0369 (MTS), 0.0210 (CTS), and 0.0129 (sex). The latter, in descending order, are: 0.1661 (MTD), 0.0475 (LTS), 0.0309 (MTS), 0.0194 (CTS), and 0.0182 (sex).

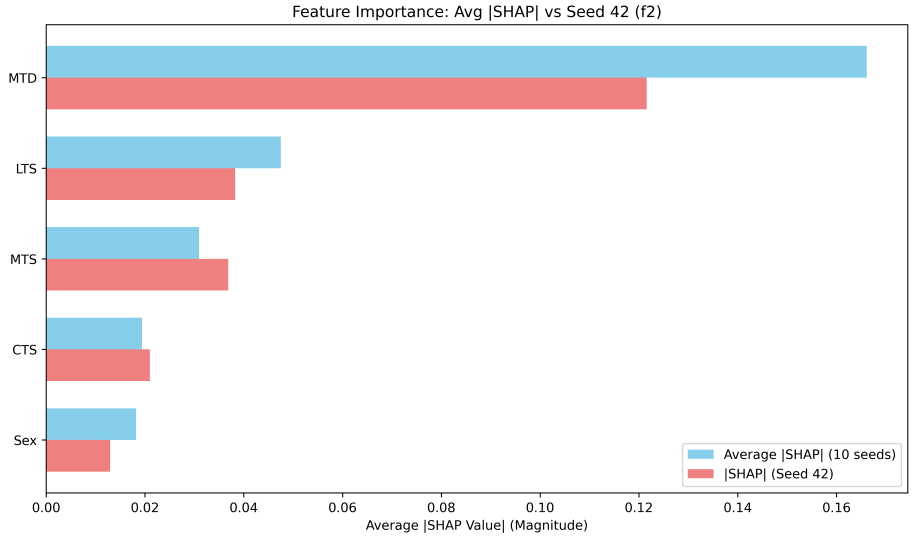


Figure 4.14: Averaged global feature importance by F2-tuned ensemble

Figure 4.15 provides clear separation of the clusters for features *MTD*, *LTS*, *MTS*, and *sex* based on their relative magnitudes. Regardless of the ordering of feature importance, decreased MTD, increased LTS and MTS, and being male are all associated with higher SHAP values contributing to predicted positives.

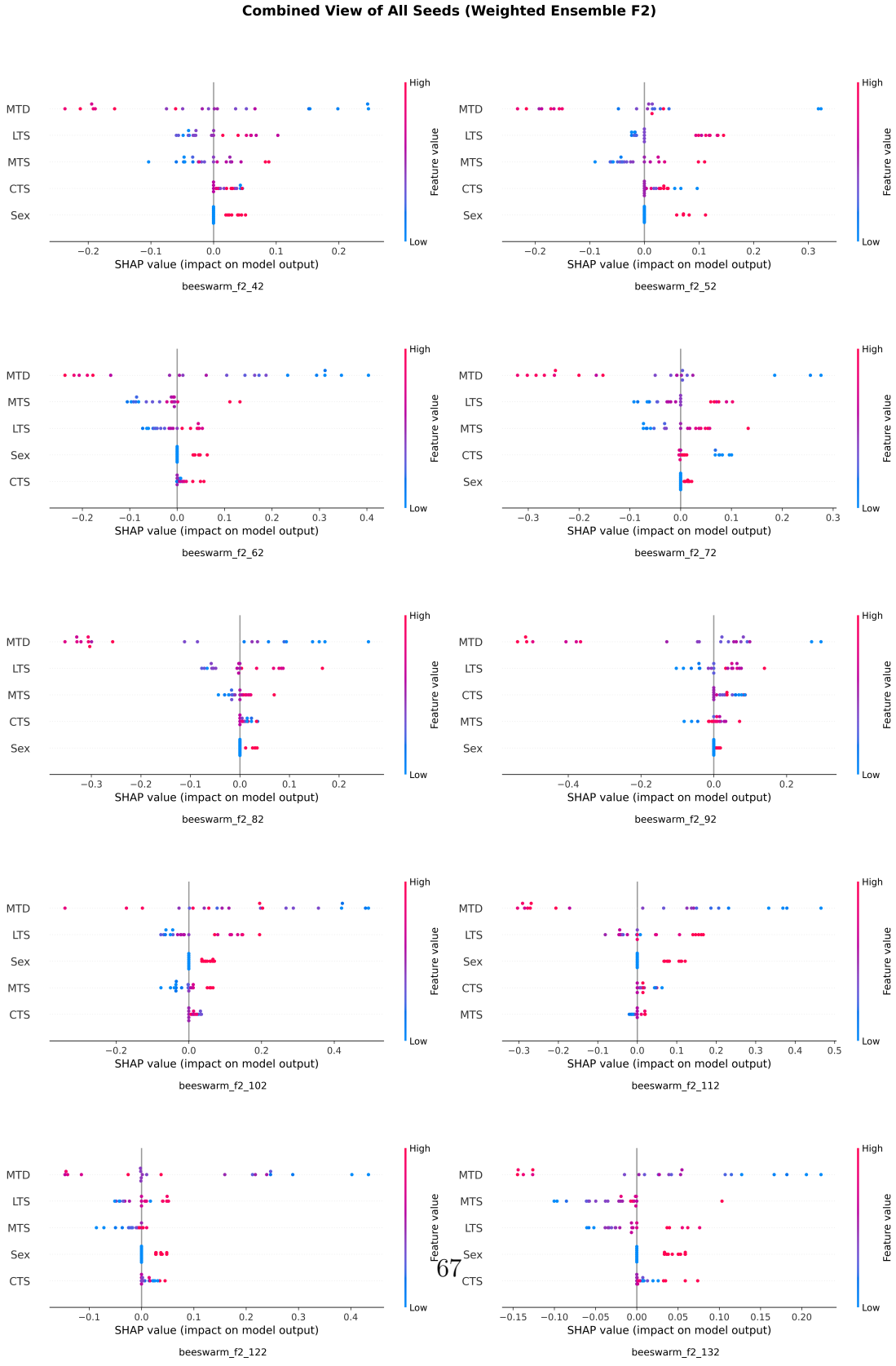


Figure 4.15: Combined view of global feature importance by F2-tuned ensemble

4.3 Full-dataset Model

The composition of the weighted ensembles are different from their standard tested counterparts. The balanced-accuracy-tuned full-dataset ensemble consists of $CatBoost\_BAG\_L1$ (weight = 0.5), $TabICL\_BAG\_L1$ (weight = 0.5). The F1-tuned full-dataset ensemble consists of $LightGBM\_r33_BAG\_L1$ (weight = 0.3333), $TabICL\_BAG\_L1$ (weight = 0.1667), $TabM\_r49\_BAG\_L1$ (weight = 0.3333), and $Mitra\_BAG\_L1$ (weight = 0.1667). The F2-tuned full-dataset ensemble consists of $LightGBM\_r33\_BAG\_L1$ (weight = 0.4286), $XGBoost\_r40\_BAG\_L1$ (weight = 0.1429), $TabM\_r49\_BAG\_L1$ (weight = 0.4286).

With the merging of test set into the train-validation set, held-out sets are no longer available for independent evaluation. As a result, this thesis provides the validation performances of these full-dataset models 4.11, and further uses the models to generate local feature explanations for comparison against their standard tested counterparts.

Table 4.11: Validation performance of full-dataset models

Model	**Validation Metric**	**Validation Score**
Full-dataset model 1	balanced accuracy	0.814
Full-dataset model 2	F1 score	0.809
Full-dataset model 3	F2 score	0.813

The local feature explanations generated by the full-dataset ensembles are plotted in Figures 4.16 to 4.18. They offer an opportunity for plausibility assessment and consistency check of the full-dataset ensembles developed for deployment.

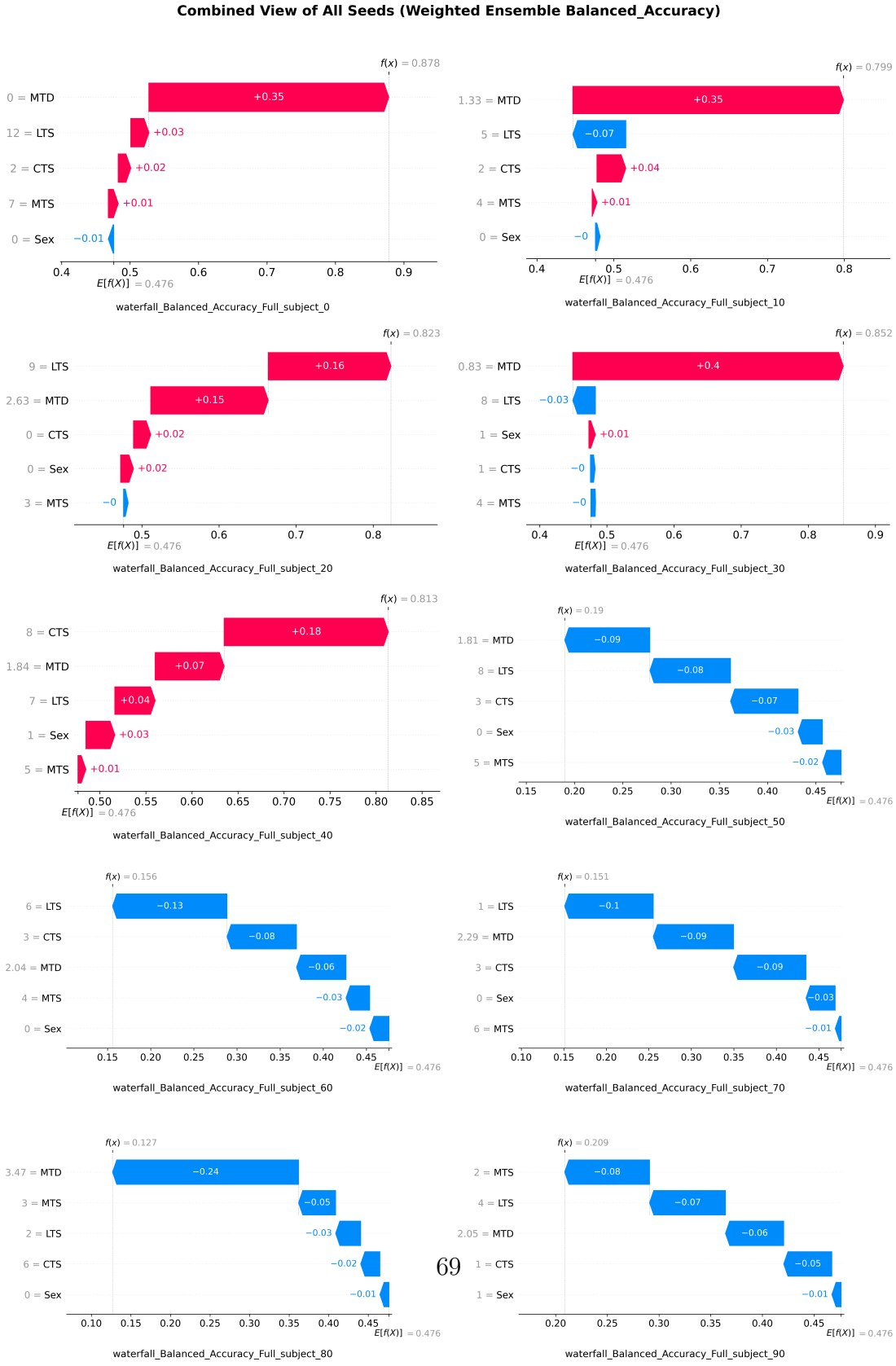


Figure 4.16: Combined view of global feature importance for balanced-accuracy-tuned ensemble

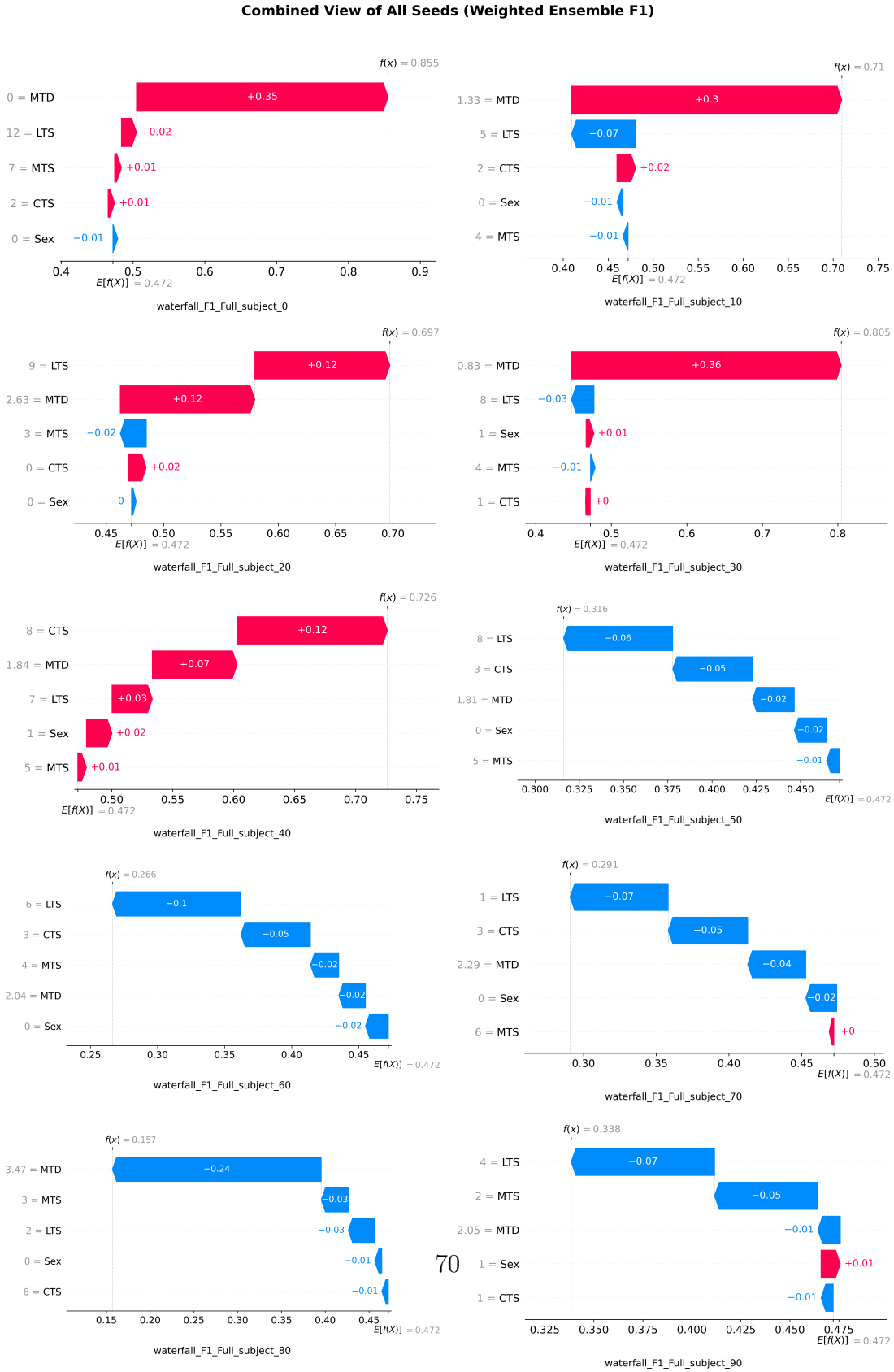


Figure 4.17: Combined view of global feature importance for F1-tuned ensemble



Figure 4.18: Combined view of global feature importance for F2-tuned ensemble

Chapter 5

Discussion

5.1 Results Summary

This study experimented with two pipelines, one developed by the author and one implemented based upon the open source library AutoGluon. Due to time constraint, the self-developed pipeline tuned the models with two different metrics (accuracy and F-1 score) while the AutoGluon pipeline tuned the models with three (balanced accuracy, F-1 score, and F-2 score). Both pipelines found numerous best-performing models, one per adopted tuning metric, resulting in five best-performing models.

The two best-performing models corresponding to the two tuning strategies from the self-developed pipeline are both TabNet models, despite TabNet being the only model type without RFE enabled in its pipeline. The instance tuned with respect to accuracy, referred to as *accuracy-tuned TabNet*, yields a validation accuracy of 0.863, test accuracy of 0.812, test F1-score of 0.800, test F2-score of 0.769, and test AUC of 0.959. The global SHAP values computed from *accuracy-tuned TabNet*, ranked in descending order, are 0.2184 for sex, 0.2043 for LTS, 0.1995 for MTD, 0.1887 for CTS, and 0.1887 for MTS, whereas the feature importance scores calculated internally by TabNet yields 0.2874 for LTS, 0.2223 for MTD, 0.1984 for CTS, 0.1711 MTS, and 0.1209 for sex. The other instance of TabNet tuned with respect to F-1 score, referred to as *F1-tuned TabNet*, yields a validation F1-score of 0.865, test accuracy 0.688, test F1-score of 0.615, test F2-score of 0.541, and Test AUC of 0.809. The global SHAP values computed from *F1-tuned TabNet*, ranked in descending order, are 0.2612 for MTS, 0.2572 for MTD, 0.1996 for CTS, 0.1947 for LTS, and 0.0869 for sex, whereas the feature importance scores calculated internally by TabNet yields 0.2612 for MTS, 0.2572 for MTD, 0.1996 for CTS, 0.1947 for LTS, and 0.0869 for sex.

The three best-performing models corresponding to the three tuning strategies from the AutoGluon pipeline are all multi-layer stack ensembles. The instance tuned with respect to balanced-accuracy, referred to as *balanced-accuracy-tuned standard ensemble*, yields a validation balanced accuracy of 0.777, test balanced accuracy of 0.950, test F1-score of 0.947, test F-2 score of 0.918, Test AUC of 0.973, test precision of 1.000, and test recall of 0.900. The global SHAP values computed from *balanced-accuracy-tuned standard ensemble*, ranked in descending order, are 0.0280 for MTD, 0.0181 for LTS, 0.0116 for MTS, 0.0068 for sex, and 0.0168 for CTS. The instance tuned with respect to F1-score, referred to as *F1-tuned standard ensemble*, yields a validation F1-score of 0.725, test balanced accuracy of 0.955, test F1-score of 0.952, test F-2 score of 0.980, Test AUC of 1.000, test precision of 0.909, and test recall of 1.000. The global SHAP values computed from *F1-tuned standard ensemble*, ranked in descending order, are 0.1216 for MTD, 0.0383 for LTS, 0.0369 for MTS, 0.0210 for CTS, and 0.0129 for sex. The instance tuned with respect to F2-score, referred to as *F2-tuned standard ensemble*, yields a validation F2-score of 0.736, test balanced accuracy of 0.955, test F1-score of 0.952, test F-2 score of 0.980, Test AUC of 1.000, test precision of 0.909, and test recall of 1.000. The global SHAP values computed from *F2-tuned standard ensemble*, ranked in descending order, are 0.0280 for MTD, 0.0383 for LTS, 0.0369 for MTS, 0.0210 for CTS, and 0.0129 for sex.

5.2 Model Choice

Due to the methodological rigor provided by AutoGluon's k-fold bagging and multi-layer stacking mechanisms, AutoGluon ensembles are presented to reader as superior options over TabNet models, despite good generalization showcased by *accuracy-tuned TabNet*. Each of the three tuning metrics adopted in this study to tune the ML models encourages distinct model behavior and properties. The balanced accuracy, for instance, assigns same weights to each mistake regardless of the class. It drives models to equally prioritize true positive rate and true negative rate, encouraging fairness between classes and making models insensitive to class imbalance in the train-validation set. The clinical outcome of selecting a balanced-accuracy-tuned model is that it will likely flag similar proportions of both classes correctly. The F1-score is relatively similar to balanced accuracy in the sense that it gracefully handles imbalanced datasets. However, instead of accounting for the true negative rate and balancing it with the true positive rate, it balances precision with recall (same as true positive rate or sensitivity) equally. This metric encourages models to focus on catching more true positives to maximize balanced positive class performance. The clinical outcome of selecting an F1-tuned model is that it will prioritize true positives while keeping false negatives and false positives reasonably controlled and balanced. The F2-

score prioritizes recall over precision, while also optimizing true positives, ensures models to favor flagging false negatives even at the cost of more falsae positives. While AutoGluon ensembles tuned with respect to all these metrics are provided, the model corresponding to the use of F2-score is the primary and the recommended one because of the cost and clinical risk associated with the misidentifying patients with high likelihood of ACL injury, as mentioned in Chapter 3.

Despite *F2-tuned standard ensemble* being the main model of interest in this study, readers and clinicians can freely select the appropriate model out of the three standard ensembles presented based on their research goal. Then, depending on the specfic use case, readers and clinicians have the freedom to choose between the standard ensemble and the full-dataset ensemble for each metric used. The standard ensemble is the safest, most conservatively evaluated version that has been independently tested on an unseen held-out set, but also suffers from higher risk of overfitting due to the extremely small dataset size. Its best use is for research benchmarking, further rigorous model integration, or relative feataure importance comparison. In contrast to the standard ensemble, the full-dataset ensemble serves as a more powerful option designed to maximize robustness and generalization for real-life clinical deployment. Despite the absence of an independent dataset used for a final unbiased evaluation, it strictly adheres to the standard train-validation paradigm which keeps it from overfitting to the train set. More discussion regarding the development and use of the full-dataset ensemble is provided below in Section 5.6.

5.3 Pipeline Comparison

The superiority of AutoGluon ensembles were assumed based on preliminary experimentation with both pipelines and the advanced, sophisticated approach proposed by AutoGluon. This assumption will be tested in this section.

5.3.1 Ground Truth Distribution on Lower-Dimensional Simplex

To delve deeper into model behavior, the ground truth distributions on the 1D simplex is plotted in Figures 5.1 to 5.5. This simplex is enclosed by the two vertices [0, 1] and [1, 0] and defined by the parametric form of $(x, y) = (t, 1 - t), t \in [0, 1]$. Geometrically, it represents the probability space of all possible distributions over the two outcome classes. On these simplexes, true negatives (blue, corresponding to actual low likelihood of ACL injury) should be located near the bottom-right vertex, whereas true positives (orange,

corresponding to actual high likelihood of ACL injury) should appear near the top-left vertex. This provides a visual means of assessing the correctness of the learned projections and the degree of separation between projected data points. The closer a projected data point is to either boundary vertices, the higher the confidence a model has in the associated predict outcome; the closer a projected data point is to the (0.5, 0.5) point on the simplex, the lower the confidence.

5.3.1.1 Self-Developed Pipeline

The ground truth simplex of *accuracy-tuned TabNet*, plotted in Figure 5.1, exhibits an extremely wide separation of data in the test set. Although all instances of 100% confidence are correctly classified by accuracy-tuned TabNet, several moderate- to high-confidence predictions are incorrect. A particular example of this is the blue data point classified with 98.06% confidence as positive (high likelihood of ACL injury). Such high-confidence mistakes undermines model trustworthiness, as clinicians cannot reliably determine prediction quality based solely on confidence scores. Ideally, models should make errors primarily on low-confidence predictions.

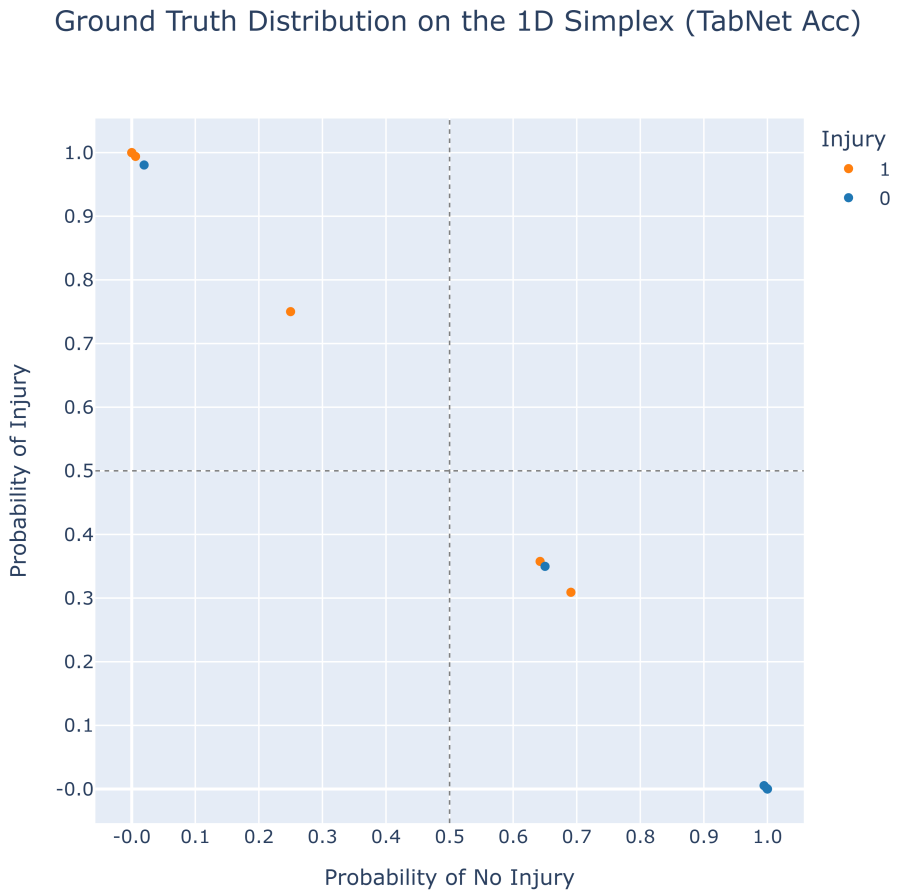


Figure 5.1: Ground truth distribution on 1D simplex for accuracy-tuned TabNet

The ground truth simplex of *F1-tuned TabNet*, plotted in Figure 5.2, also shows extremely wide separation of data in the test set. Similar to *accuracy-tuned TabNet*, *F1-tuned TabNet* misclassifies a true positive with high confidence of 90.91% as negative (low likelihood of ACL injury). Moreover, both boundary vertices of the simplex contain instances from both outcome classes. Specifically, the projected data points lying on the (0, 1) vertex, corresponding to 100% confidence in the positive class, consist of two true positives and one true negative. On the other hand, the projected data points lying on the (1, 0) vertex, corresponding to 100% confidence in the negative class, comprise three actual positives and seven actual negatives. This makes *F1-tuned TabNet* a worse prognostic model than *accuracy-tuned TabNet*.

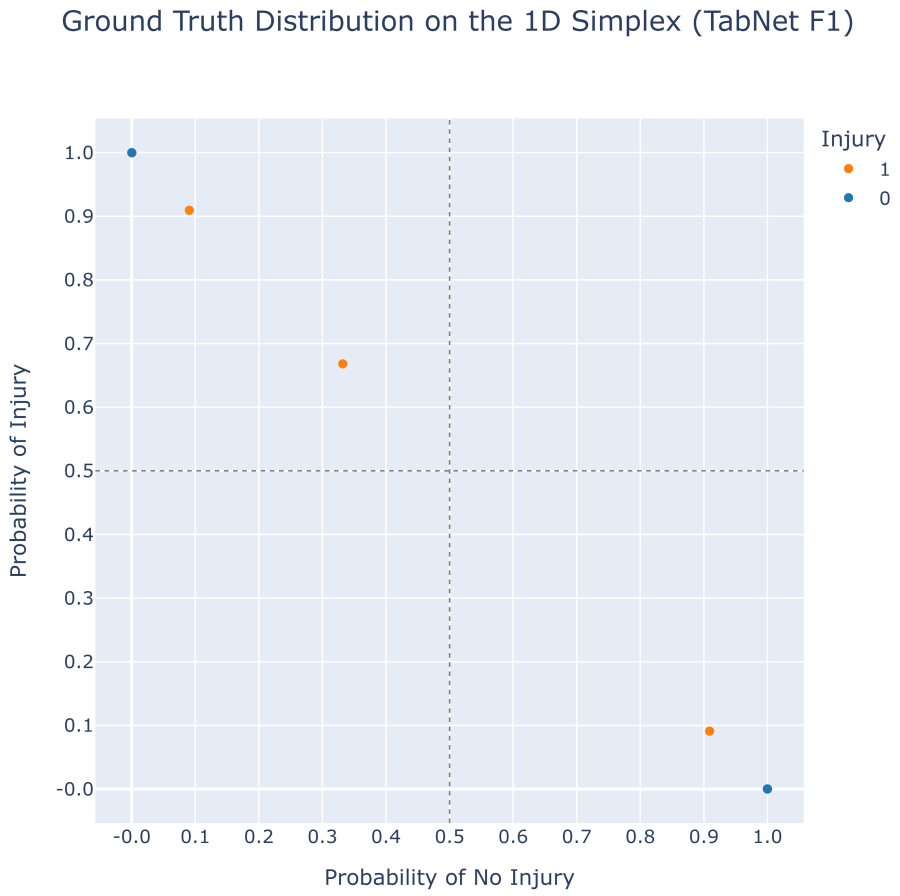


Figure 5.2: Ground truth distribution on 1D simplex for F1-tuned TabNet

Despite decent generalization to the validation subsets, TabNets are not stable when performing inference on the test set. This includes *accuracy-tuned TabNet* which, despite achieving high predictive scores, produces unreliable confidence estimates. As shown in the ground truth simplex visualizations, both TabNet models produce overconfident predictions in both directions, even when the corresponding outcomes are mistaken. This behavior arises from TabNet's use of the *entmax* and *sparsemax* masking functions proposed by [6]. While these sparse normalizations sparsify the low-confidence region of the probability space to improve feature selection effectiveness, it compromises the model's ability to give meaningful confidence estimates. Regardless of correctness, most data points are projected toward one of the boundary vertices, rendering the confidence scores uninformative. Such extreme confidence in erroneous predictions, also documented in other modern DNNs [105,

93, 113], signifies the TabNet models have adopted decision boundaries that do not truly generalize and remain overly confident on OOD data. The lack of epistemic awareness limits their suitability for feature importance analyses and contributes to the feature importance explanation instability which will be discussed in Section 5.3.2.

5.3.1.2 AutoGluon Pipeline

The ground truth distribution of *balanced-accuracy-tuned standard ensemble*, plotted in Figure 5.3, shows characteristic almost opposite of what was observed for TabNet models. For *balanced-accuracy-tuned standard ensemble*, all instances in the test set are transformed and squeezed into the (0.4, 0.6) range in the probability space. This compression reflects the model's conservative prediction behavior, producing moderate confidence scores rather than extreme ones. Such behavior aligning with the comparatively small SHAP values computed from the model tabulated in Table 4.10. This level of separation, though usable, is not ideal for clinical deployment since it does not meaningfully communicate the model's confidence in predictions made on unseen data.

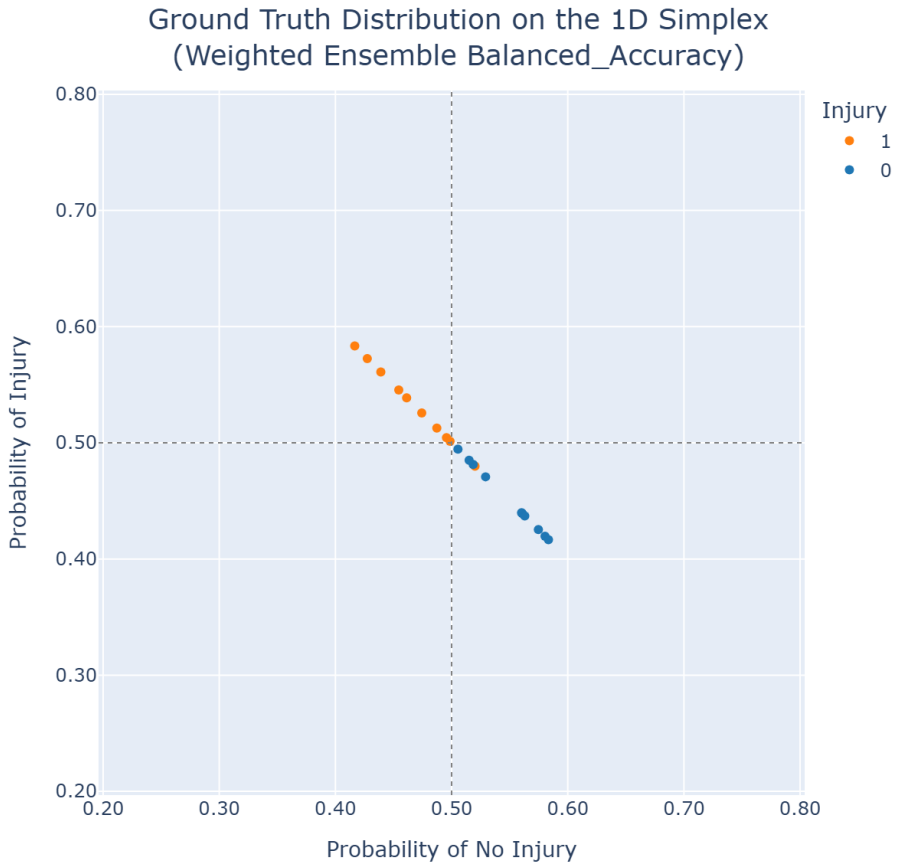


Figure 5.3: Ground truth distribution on 1D simplex for balanced-accuracy-tuned ensemble

The ground truth distribution of *F1-tuned standard ensemble* and *F2-tuned standard ensemble*, as shown in Figures 5.4 and 5.5, occupy a highly desirable region of the probability space, with all predictions embedded with in the [0.2, 0.8] range. This range represents an effective compromise between the extreme overconfidence exhibited by the TabNet models and under-confidence by *balanced-accuracy-tuned standard ensemble*. Furthermore, both *F1-tuned standard ensemble* and *F2-tuned standard ensemble* make mistakes only on data points of low prediction confidence, as demonstrated by the blue point with 53.19% confidence in likelihood of injury. The well-spread and controlled probabilities make these two ensembles particularly well-suited for clinical deployment and reliable feature attribution.

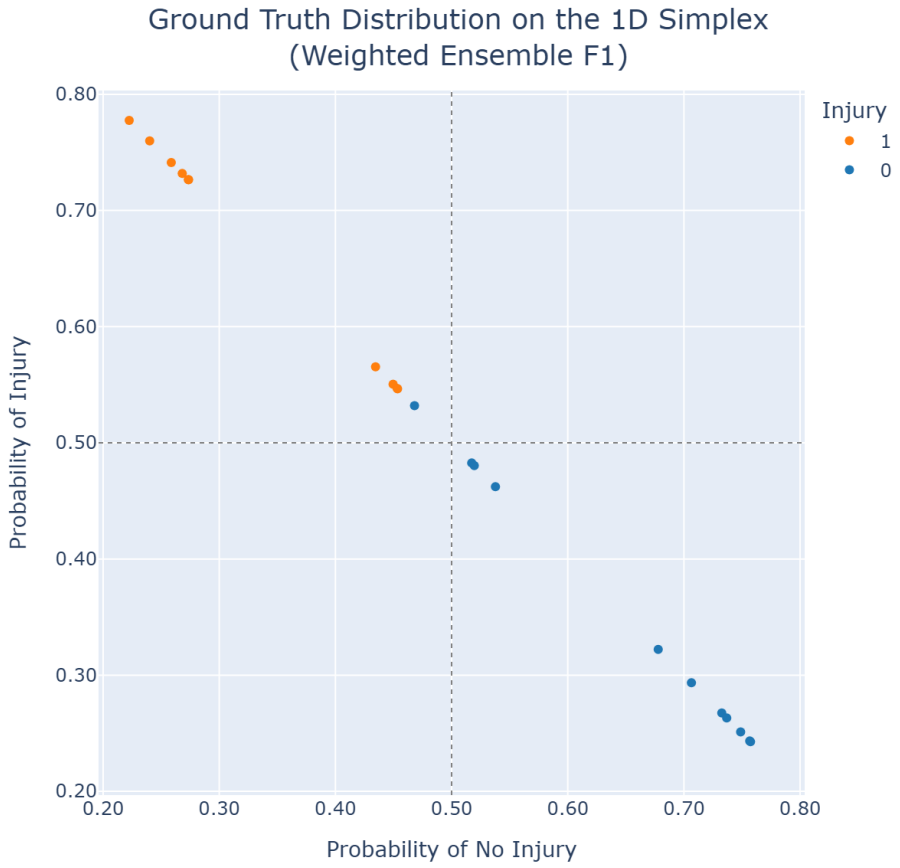


Figure 5.4: Ground truth distribution on 1D simplex for F1-tuned ensemble

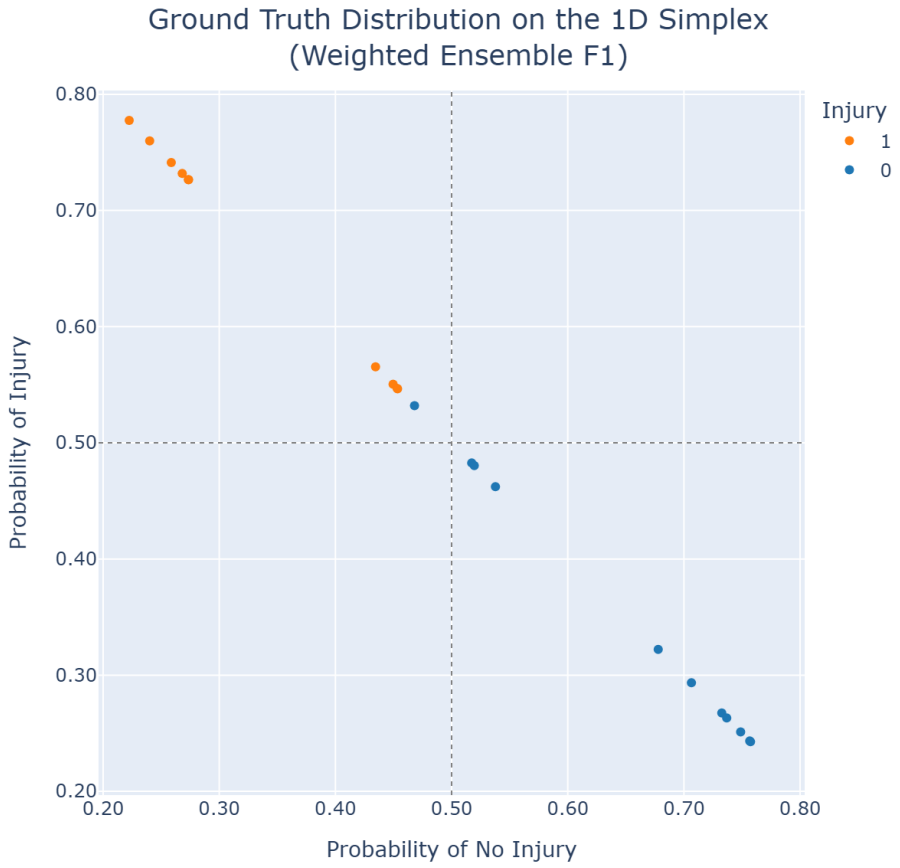


Figure 5.5: Ground truth distribution on 1D simplex for F1-tuned ensemble

In contrast to TabNet's systematic overconfidence, *balanced-accuracy-tuned standard ensemble* showed the opposite extreme behavior, conservatism, highlighting a different but equally problematic failure mode in uncertainty quantification. With all data compressed into a low-probability region, the model devises a prediction strategy that avoids strong confidence even when correct. While this prevents high-confidence errors, it severely limits the model's ability to communicate meaningful uncertainty, rendering confidence scores largely uninformative for clinical decision-making as well. *F1-tuned standard ensemble* and *F2-tuned standard ensemble* strike an optimal balance between the previous models, demonstrating safe and well-calibrated failure modes for reliable feature importance and analyses.

5.3.2 Global Feature Explanation Stability

The best-performing models from both pipelines are assessed based on the stability of their generated global feature explanations.

5.3.2.1 Self-Developed Pipeline

For TabNet models obtained from the self-developed pipeline, global feature importance can be calculated with both SHAP values and the summation of TabNet's generated masks. While the study focuses on SHAP due to its (I) theoretical soundness, (II) directionality in explanation, and (III) model agnostic nature, comparing the outcome of the two approaches can provide some supplementary insights.

The computed global SHAP values are normalized for fair comparison against TabNet's internal global feature importance scores. The normalized comparison for *accuracy-tuned TabNet* and *F1-tuned TabNet* can be seen in Figures 5.6 and 5.7, respectively. The global feature explanations for the accuracy-tuned TabNet largely align (Figure 5.6). Both feature importance calculation methods assigned incrementally lower scores to LTS, MTS, CTS, and MTS in that order. However, SHAP ranks sex as the most informative feature while TabNet ranks it as the least informative. The global feature explanations for *F1-tuned TabNet* (Figure 5.7), on the other hand, show no consistent patterns across the two method besides ranking sex as the least informative features.

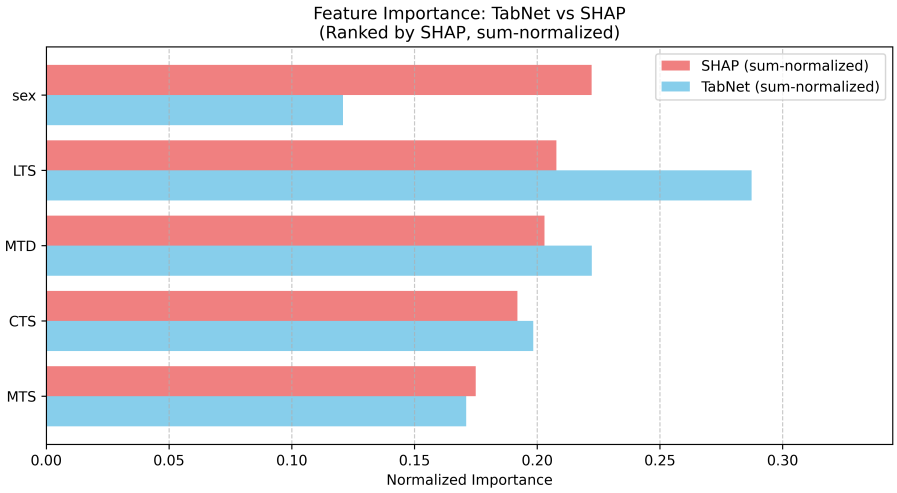


Figure 5.6: Feature importance comparison for accuracy-tuned TabNet

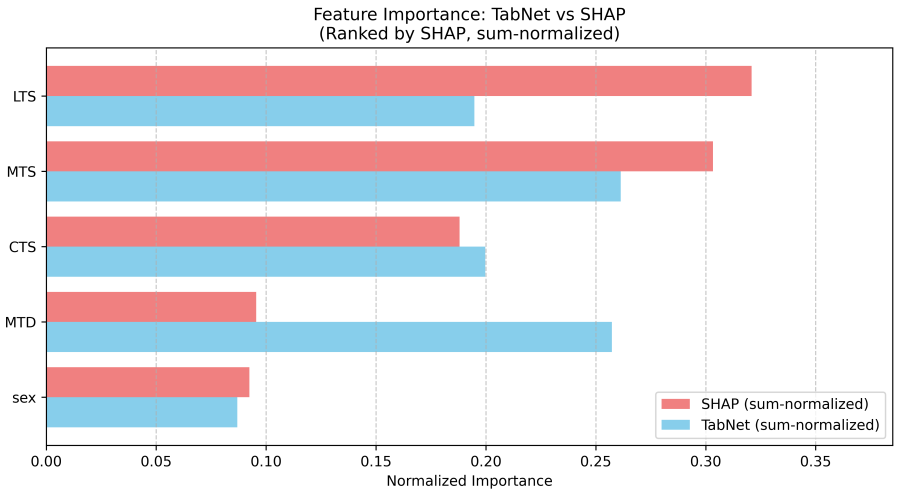


Figure 5.7: Feature importance comparison for F1-tuned TabNet

Then, comparing the global SHAP values across the two models in Figure 5.8, it is obvious that the results largely diverged. The difference in patterns indicate that TabNet is highly sensitive to the tuning metrics when applied to the provided dataset, which was previously implied by the disparity in raw performance between Tables 4.1 and 4.3. While different patterns is common for models tuned with different metrics, such inconsistency is too significant to be justifiable. Findings in Section 5.3.1 points to similar conclusions for TabNet models. Taking the mean between the two models for a more stable TAF importance ranking, Figure 5.9 can be obtained, yields LTs as the most contributive feature and MTS as the least.

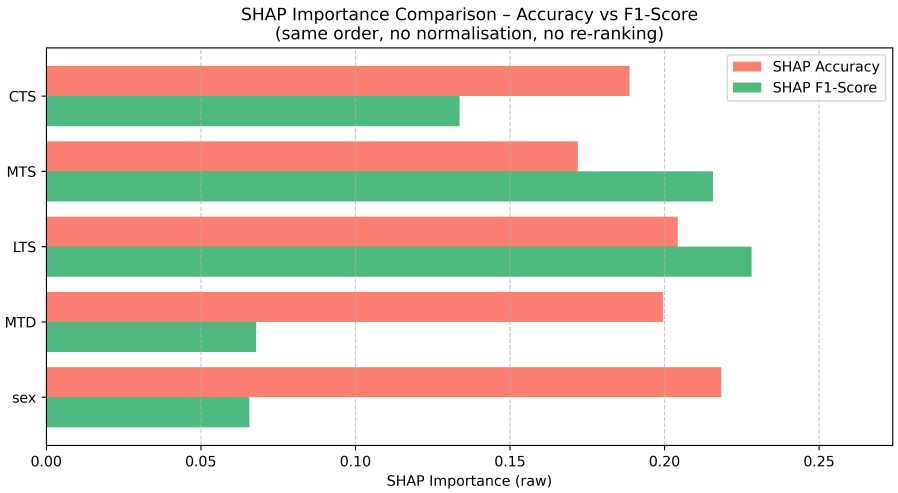


Figure 5.8: Comparison of global SHAP values across two TabNet models

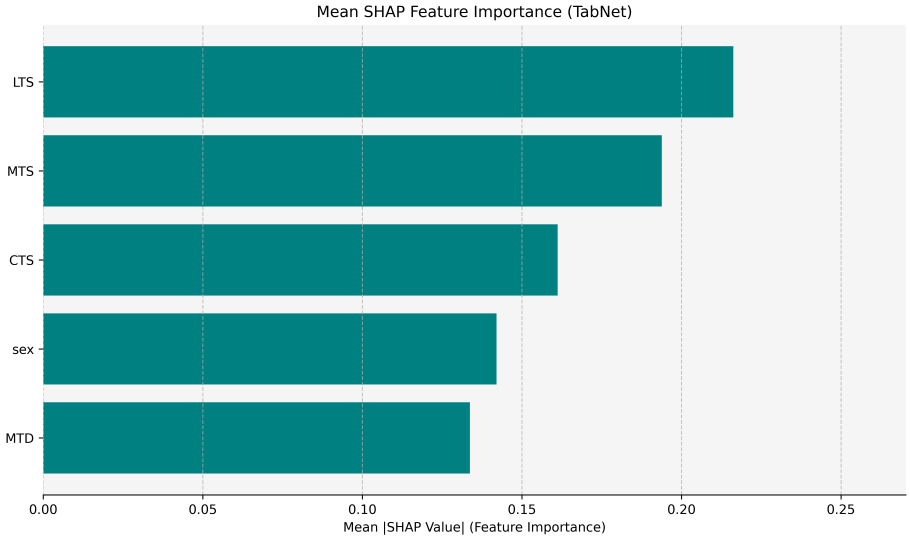


Figure 5.9: Mean global SHAP values across tuning metrics for more stable ranking

Despite improved stability via feature importance aggregation, results from 5.9 are presented mainly for completeness and not used for subsequent feature importance quantification. This exclusion is deliberate. Taking into consideration SHAP values are computed as differences in prediction confidence scores output by the model across different feature

coalitions, and that confidence scores generated by the TabNet models are flawed and unreliable as discussed in Section 2.1, the resulting SHAP values do not faithfully represent true causal or predictive contributions. Any clinical or biomechanical conclusions drawn from the two TabNet models risk being misleading and biased.

5.3.2.2 AutoGluon Pipeline

The global feature explanations computed from best-performing AutoGluon ensemble models obtained are readily compared against the average global feature explanations calculated across all ten seeds (Figures 4.10 to 4.14). For the balanced-accuracy-tuned ensembles, the order of feature importance on the default seed and the ten-seed-aggregation fully aligned, ranked in the order of MTD, LTS, MTS, sex, and CTS. For the F1-tuned ensembles, the order of feature importance are mostly consistently. For the F1-tuned ensembles, both the default seed and the ten-seed-aggregation gave MTD, LTS, and MTS as the three most contributive TAFs. The only difference lies in the ranking of CTS and sex, but their respective SHAP values are different by a small margin. For the F2-tuned ensembles, the order of feature importance are fully consistent. the order of feature importance on the default seed and the ten-seed-aggregation are fully aligned, ranked in the order of MTD, LTS, MTS, CTS, and sex.

Comparing the global SHAP values across the three best-performing AutoGluon ensemble models in Figure 5.10, global feature explanations nearly fully aligned, with minor differences in rankings of CTS and sex. It is worth mentioning that the small SHAP values from *balanced-accuracy-tuned standard ensemble* correspond to the low confidence scores generated by that model. The same robust ranking calculated by averaging SHAP values across different metric is done in Figure 5.11, similar to for that of TabNet. However, considering that *balanced-accuracy-tuned standard ensemble* outputs overly conservative predictions, *F1-tuned standard ensemble* and *F2-tuned standard ensemble* are prioritized for feature attribution.

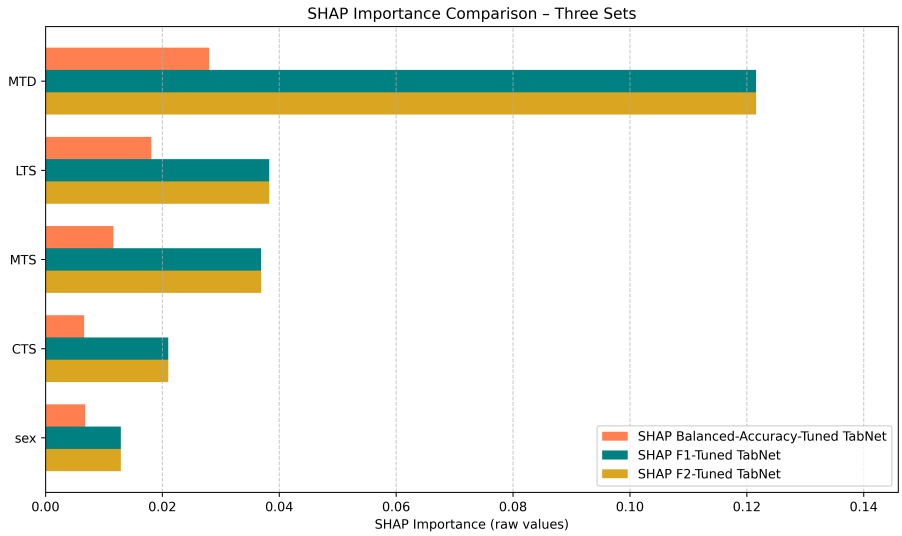


Figure 5.10: Comparison of global SHAP values across three ensemble models

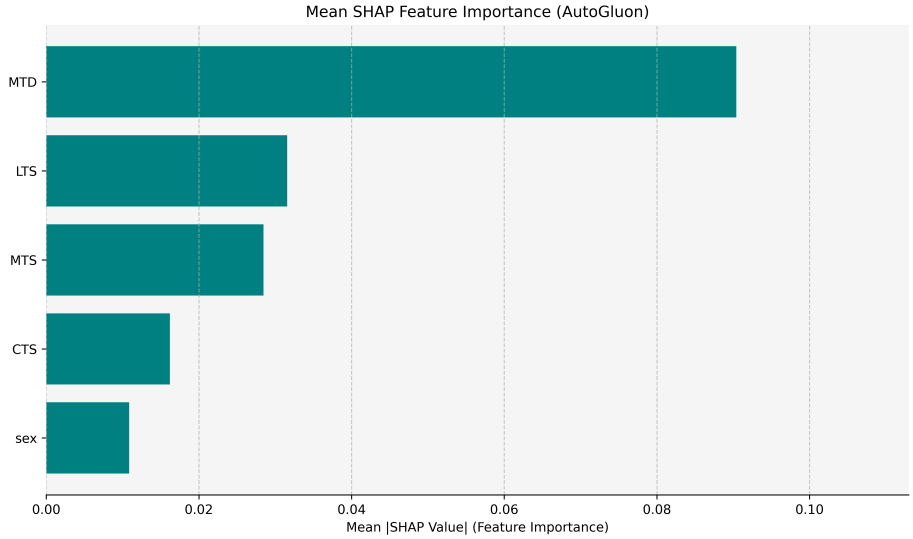


Figure 5.11: Mean global SHAP values across tuning metrics for more stable ranking

5.4 Literature Comparison

5.4.1 Predictive Performance

To the best of the author's knowledge, the study by Tamimi et al. represents the only prediction-focused study among all literature that utilized ML as a tool for predicting the likelihood of ACL injury using the different components of posterior tibial slope [108]. This makes the model developed by Tamimi et al. the only appropriate, direct benchmark for comparison in the present study. Comparing the predictive power of the recommended *F2-tuned standard ensemble* presented by this study (balanced accuracy: 95.5%, accuracy: 95.2%, sensitivity: 100.0%, and specificity: 90.9%) against that of Tamimi et al.'s (accuracy: 92%, sensitivity: 92%, and specificity: 92%), performance is highly comparable. Notably, both models yield optimistic estimates due to favored data splits, as indicated by the discrepancy between validation and test scores. Realistically, the models' performance on other sets of unseen independent data is anticipated to be closer to the validation scores. Furthermore, the resemblance in validation and test performance between the two studies signifies that the explainable SOTA models employed in this study do not provide any meaningful improvement over Tamimi et al.'s implementation of Gaussian naive Bayes model, despite the fact that they are designed to tackle small tabular datasets and take into account nonlinear feature interactions. This suggests the problem is data-limited rather than model-limited. Some examples of causes are the small datasets used and limited set of features included, which will be discussed in Chapter 7. Despite accomplishing no observable improvement on model performance, the current study validates the predictive utility of anatomical features and further contributes to clinical injury likelihood assessment by delving into model characteristics (Section 5.3.1 and detailed feature importance through SHAP (Sections 5.3.2 and 5.5), which is lacking in [108].

5.4.2 Feature Importance and Potential Risk Factors

Global SHAP analysis was used to elucidate the most contributive tibial TAFs from the recommended *F2-tuned standard ensemble*. Globally, the *F2-tuned standard ensemble* and the ten-seed aggregation concurred, both concluding that feature importance is ranked in the order of MTD, LTS, MTS, CTS, and sex, despite slight differences in numerical values. Further, a closer investigation of the beeswarm plot suggests a decreased MTD, increased LTS, and increased MTS as top predictors contributing to higher predicted likelihood of ACL injury.

The findings regarding feature importance from the presented models mostly align

with existing literature. For instance, increased LTS and MTS have been identified by review studies as important risk factors of non-contact ACL injury, influencing ACL force distribution, translation on the lateral tibial plateau, and pivot shifts to the medial tibial plateau [10, 125, 114]. Similarly, [88] has summarized across studies, that both increased LTS and MTS are widely regarded as risk factors, particularly in females, where they may exacerbate injury likelihood. Furthermore, increased MTS and LTS have been established by a review study as risk factors for ACL injury, with the ratio of MTS to LTS showing a negative correlation with knee abduction, consistent with current understandings of ACL injury mechanisms [23]. Despite slight difference in results due to geographical variation, it is becoming increasingly evident that LTS and MTS are crucial for predicting ACL injury, and that slope-reducing osteotomy should be further investigated as a potential prevention mechanism [125]. As an adjacent to other previously discussed tibial slopes, CTS is identified as a weak predictor of ACL injury. This finding concurs with the limited number of available literature on CTS, which concluded that CTS was not a risk factor [76, 41].

Unlike tibial slopes, the MTD is much more complicated. On one hand, studies like [41, 89] identified MTD as a risk factor. On the other hand, some works failed to observe statistical significance in the difference between MTD across groups [66, 45]. This likely stems from the fact that MTD is not an independent risk factor, which makes it more difficult to identify using traditional statistical tests. Biomechanically, LTS and MTS directly and independently contribute to the forces exerted on the tibia and, in turn, the ACL. MTD, on the contrary, does not have a discernible effect until combined with a steep tibial slope [88]. The depth of the tibial plateau determines the geometric constraints on knee kinematics; without a larger MTD to interfere with the translation induced by increased LTS and MTS, the posterior translation of the tibia is unrestrained, placing increased strain on the ACL and elevating the risk of rupture [102]. Therefore, MTD is a dependent but synergistically critical predictor of ACL injury, yielding the highest global SHAP value in this study.

Sex being elucidated as the least predictive feature was the most surprising result for the authors, given that sex is commonly recognized as a risk factor of ACL injury due to the higher incidence observed among females [14]. This unanticipated result is attributed to sex being correlated with other variables and features. Some works, for example, categorize sex as an intermediate risk factor, suggesting its association with ACL injury is mediated by tibial slopes, MTD, and related dynamic instability in the tibiofemoral region [42, 53, 25]. According to these studies, women were observed to have increased LTS and MTS with decreased MTD, which aligns with Table 3.1. It is clear from Table 3.1 that female participants consistently have higher MTS and LTS as well as lower

MTD when compared against their male counterparts. As a result, the presented models attend more to the TAFs which contain richer semantics.

While clinical thresholds could be derived, the authors argue against their use since they do not adequately capture the complex non-linear interactions between different TAFs. For instance, the knee kinematics as a function of MTS, LTS, and MTD could be highly convoluted. The same numerical value of MTD may correspond to a positive or negative instance, depending on the magnitude of the MTS and LTS. Two examples of this are showcased by *subject*_20 and *subject*_80 in Figures 5.12 to 5.14, where the instances share the same MTS value, but have different predicted outcomes (and ground truth labels). Instead of relying on clinical thresholds, the integration of explainable ML models can offer superior performance. A more comprehensive analysis regarding cut-off thresholds for feature importance, if necessary, can be conducted by examining the distributions in SHAP stacked force plots.

5.5 Local Feature Importance

Local feature importance are enabled to facilitate decision making in real-life clinical applications, providing insights into the model's decision on individual patients. Combined views of examples are presented as SHAP waterfall plots in Figures 5.12 to 5.14. The expected value $E[f(X)]$ in each plot represents the prediction confidence score when a model is provided as inputs feature-wise medians. The final value $f(X)$ is the final predicted probability of ACl injury for an individual patient given their personal TAFs, corresponding to the y-value on the 1D simplexes shown in Section 5.3.1. The arrows and colors indicate the magnitudes and directions in which different TAFs and sex drive model predictions. Of all 104 data instances present in the dataset, ten (selected by increments) are plotted for demonstration. It can be seen from the waterfall plots that the local feature explanations generated by the three models are highly similar and numerically consistent. The consistency validates the reliability of the ensemble models and boosts the trustworthiness of the models' learned representations and feature importance rankings.

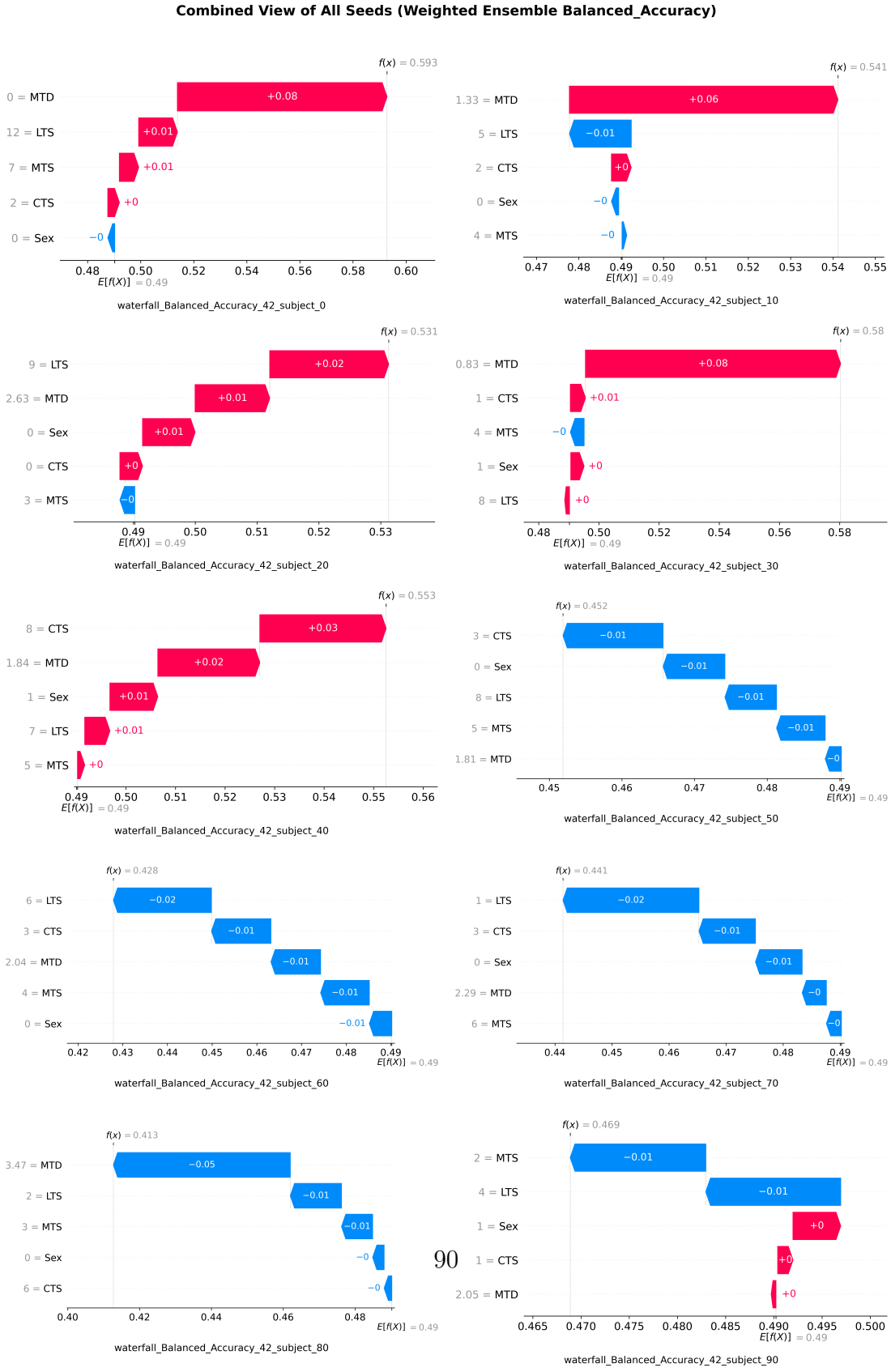


Figure 5.12: Combined view of global feature importance for balanced-accuracy-tuned ensemble



Figure 5.13: Combined view of global feature importance for F1-tuned ensemble



Figure 5.14: Combined view of global feature importance for F2-tuned ensemble

5.6 Full-Dataset Ensemble

The default seed of 42 appears to provide an unusually favorable split. Across all three tuning metrics, models trained on this split achieved the highest test performance among the seeds examined, consistently outperforming their corresponding validation score by 0.173 to 0.244. This systematic validation-test discrepancy indicates optimistically biased evaluation results.

Taking into consideration high sensitivity of results to random splitting in small datasets, full-dataset ensemble alternatives are developed for each of the AutoGluon standard ensembles. The full-dataset ensembles are trained on the entire available dataset by merging the original training, validation, and test sets, thereby ensuring that every observation contributes equally to learning. All training and tuning techniques, including k-fold bagging, validation, early stopping, remain strictly adhered to the gold standard of ML practice. Model selection, however, is restricted to the best-performing model type as it shows the best chance of constructing a successful learner. Inheritance of specific model architectures and hyperparameters are avoided to prevent overfitting to the original train-val set. There is no final evaluation as full-dataset models are designed solely for robust and generalizable deployment.

This approach is argued to be justifiable because this thesis does not conduct FIA or claim superior model performance by evaluating the full-dataset ensembles on any portion of the data, but merely proposes the use of them for future clinical deployment as they are better constrained by data and are statistically less likely to make mistakes in comparison to their standard ensemble counterparts. Considering the extremely small size of the dataset, there is hardly any guarantee that even the full dataset is representative of the true distribution in the real world. Given that the test models which have been thoroughly examined are transparent to readers and used as foundations to build the final full-dataset ensembles, this approach is deemed a necessary and justified compromise to make the models as robust as possible.

This argument is backed up by the local feature importance which expresses almost identical patterns to their standard ensemble counterparts. The only significant differences in the local feature explanations from the two training paradigms are the numerical output values $f(X)$. Full-dataset ensembles produce larger output values, showing that they are able to utilize the additional data from the original test set to further push the two classes apart in the probability space through learned feature representations. The balanced-accuracy-, F1- and F2-tuned full-dataset ensembles achieves validation scores of 0.814, 0.809, and 0.813, respectively. Additionally, their local feature explanations remain consistent across tuning metrics, and are similar to the ones obtained from the standard

ensembles.

5.7 Model Suitability

5.7.1 Self-Developed Pipeline

Table 5.1: Best and average F1-tuned model ranking

Model	**Rank**
TabNet	1
CatBoost	2
XGBoost	3
RF	4
SVM	5
KNN	6
MLP	7

Table 5.2: Best and average F1-tuned model ranking

Model	**Rank**
TabNet	1
CatBoost	2
XGBoost	3
RF	4
MLP	5
SVM	6
KNN	7

Based on the validation-score-ranked Tables 5.1 and 5.2, TabNet appears to be the best option among the pool of models for the designated small tabular dataset binary classification task. The next best options are CatBoost, XGBoost, and RF, which are all GBTs. Finally, the worst model choices are SVM, KNN, and MLP.

This confirms the trends documented in literature: (I) GBTs outperform most other common alternative approaches, and (II) specialized designs of DNNs such as TabNet can be the best model choice for small tabular datasets, outperforming even GBTs.

5.7.2 AutoGluon Pipeline

Table 5.3: Best and average balanced-accuracy-tuned model ranking

Model	Best Rank	Average Rank
Weighted Ensemble	1	1.00
TabM	2	8.67
CatBoost	4	6.00
LightGBM	10	15.33
XGBoost	13	14.50
MITRA	17	17.00
TabICL	18	18.00
TabPFNv2	19	20.00

Table 5.4: Best and average F1-tuned model ranking

Model	Best Rank	Average Rank
Weighted Ensemble	1	1.00
TabM	2	5.00
LightGBM	7	12.33
XGBoost	9	11.50
MITRA	11	11.00
CatBoost	13	15.20
TabICL	17	17.00
TabPFNv2	19	20.00

Table 5.5: Best and average F2-tuned model ranking

Model	Best Rank	Average Rank
Weighted Ensemble	1	1.00
TabM	2	5.33
LightGBM	7	12.67
XGBoost	8	13.00
CatBoost	10	13.60
MITRA	11	11.00
TabICL	17	17.00
TabPFNv2	19	20.00

Based on the validation-score-ranked Tables 5.1 to 5.5, AutoGluon's Layer 2 (L2) weighted ensembles appear to be the best option among the pool of models for the designated small

tabular dataset binary classification task. The next best option is TabM, followed by GBTs, and eventually ICL-based DNNs. This demonstrates that GBTs are stronger candidates for such task than ICL-based DNNs, but fall short to specialized non-ICL DNNs, confirming the same trends.

While this produces an hierarchy of specialized non-ICL DNNs, GBTs, and ICL-based DNNs, AutoGluon is able to reliably achieve even better performance than any of the individual components in the hierarchy through multi-layer stacking. For instance, the balanced-accuracy-tuned version, ensembled $CatBoost\_r24\_BAG\_L1$ (5th), $TabM\_r184\_BAG\_L1$ (12th), and $XGBoost\_r40\_BAG\_L1$ (13th) to improve the validation performance by 0.013 and test performance by 0.004; the F1-tuned version ensembled $TabM\_r49\_BAG\_L1$ (3rd) and $LightGBM\_r11\_BAG\_L1$ (7th) to improve validation performance by 0.010 and test performance by 0.008; the F2-tuned version ensembled $TabM\_r49\_BAG\_L1$ (2nd) and $LightGBM\_r11\_BAG\_L1$ (9th) to improve validation performance by 0.007 and test performance by 0.010. Interestingly, every standard ensemble model presented here consists of models from both the GBT family and the DNN family.

Table 5.6: Best and average balanced-accuracy-tuned full-dataset model ranking

Model	**Best Rank**	**Average Rank**
Weighted Ensemble	1	1.00
CatBoost	2	8.00
LightGBM	3	13.33
TabM	4	10.33
TabICL	5	18.00
XGBoost	12	14.50
MITRA	16	16.00
TabPFNv2	19	20.00

Table 5.7: Best and average F1-tuned full-dataset model ranking

Model	**Best Rank**	**Average Rank**
Weighted Ensemble	1	1.00
TabM	2	7.17
LightGBM	3	12.67
TabICL	4	4.00
XGBoost	8	13.00
MITRA	10	10.00
CatBoost	11	14.20
TabPFNv2	19	20.00

Table 5.8: Best and average F2-tuned full-dataset model ranking

Model	**Best Rank**	**Average Rank**
Weighted Ensemble	1	1.00
TabM	2	7.67
TabICL	4	4.00
LightGBM	5	15.33
CatBoost	6	11.40
MITRA	11	11.00
XGBoost	12	16.50
TabPFNv2	17	18.33

For the full-dataset ensemble models, the trends previously observed becomes less obvious but still holds.

The balanced-accuracy-tuned full-dataset ensemble consists of $CatBoost\_BAG\_L1$ (2nd) and $TabICL\_BAG\_L1$ (5th); the F1-tuned full-dataset ensemble consists of $TabM\_r49\_BAG\_L1$ (2nd), $LightGBM\_r33_BAG\_L1$ (3rd), $TabICL\_BAG\_L1$ (4th), and $Mitra\_BAG\_L1$ (10th); the F2-tuned full-dataset ensemble consists of $TabM\_r49\_BAG\_L1$ (2nd), $LightGBM\_r33\_BAG\_L1$ (5th), and $XGBoost\_r40\_BAG\_L1$ (12th). Similar to standard ensembles, all full-dataset ensembles constructed also consist of models from both the GBT and DNN families.

This section verifies from the self-developed pipeline that the order of model suitability for this task, in descending order, is tabular-data-specific (self-supervised) DNNs, GBTs, and then the rest which includes instance-based methods, kernel-based methods, and generic DNNs. It then verifies from the AutoGluon pipeline that the order of model suitability for this task, in descending order, is L2 weighted ensembles, tabular-data-specific DNNs, GBTs, and then ICL-based DNNs. Both standard ensembles and full-dataset ensembles proved that performance gains can be effectively achieved by leveraging combined strengths from different model families.

Chapter 6

Conclusion

In conclusion, this study presents six explainable SOTA prognostic models for predicting ACL injury based on TAFs and sex, two of which are recommended (one standard ensemble model and one full-dataset ensemble model) based on the recommended tuning metric. These models achieved moderately high performance. The recommended standard ensemble model, *F2-tuned standard ensemble*, achieved a validation F2-score of 0.736 as well as test balanced accuracy of 0.955, test F1-score of 0.952, test F2-score of 0.980, test ROC AUC of 1.000, test precision of 0.909, and test recall of 1.000. The recommended full-dataset model, *F2-tuned full-dataset ensemble*, achieved a validation F2-score of 0.813. These demonstrate a satisfactory level of predictiveness for the models. The global SHAP explanations obtained from the recommended models aligned with existing literature, identifying decreased MTD, increased LTS, and increased MTS as the most contributive tibial plateau morphological features. Sex and CTS did not have substantial effects on the predictions. Confidence scores and local SHAP values were implemented to offer clinicians and other users insights into the decision process behind the black-box model. Overall, the models developed through this study laid the foundation for future research and can be further integrated with existing ACL registries to improve generalizability or combined with other studies to accomplish more complex tasks.

Chapter 7

Limitation and Future Work

Limitations inherent in this study are acknowledged, including dataset size, feature availability and quality, and uncertainty quantification.

Firstly, the dataset utilized is considerably smaller than that required for the problem to be Probably Approximately Correct learnable [43]. In the absence of this property, a significant risk exists that the data sampled from the underlying distribution may be unrepresentative, rendering the model prone to overfitting. Consequently, the model's capacity to generalize effectively to unseen data and to draw robust conclusions regarding the predictive power of features is compromised. Ideally, increasing the dataset size would serve as a superior regularization strategy that could result in better convergence of model predictions and feature explanations, as extensively proven from the perspective of VC-dimension [109], bias-variance tradeoff [34], or empirical evidence [9, 48]. In fact, the whole premise of data augmentation is to strengthen constraints on the hypotheses class by providing more data points, as shown in [69, 47, 117]. However, this is not always feasible for ML studies in orthopaedics due to cost and complexity constraints associated with large-scale data collection and annotation.

Some other limitations include the number of features available and the quality of them. The TAFs currently use, as pointed out by [116, 125], still have no agreed-upon consensus thus far in literature. Both the *Hashemi* method and the *Hudek* methods are commonly seen in the literature, each having their advantages and disadvantages. Without a gold standard for measurement techniques, different research groups would generate semantically heterogeneous datasets, which in turn confuses ML models and undermines reproducibility. Furthermore, the feature set employed in this thesis was inherently limited. Specifically, this thesis incorporated only a subset of biomechanical features derivable from MRI scans, such as tibial slopes and depths, while omitting other potentially rele-

vant anatomical factors like cartilage thickness, femoral notch width, ACL volume, and meniscal slopes [125, 115, 10] due to data unavailability. Beyond time-invariant anatomical features, temporal features such as age, time of injury, and other factors are also largely absent. These factors are likely essential for establishing a precise mapping between input features and the target outcome of ACL injury. Operating within a constrained and partially standardized feature space forces models to learn highly complex, conditional probability distributions rather than simpler, more deterministic relationships. This inherently introduces aleatoric uncertainty and ambiguity to the learning task, potentially causing performance plateaus observed in the SOTA models implemented.

Finally, the models currently developed do not posses the ability to differentiate between aleatoric and epistemic uncertainties based solely on their confidence score. That is, the models cannot distinguish the uncertainty due to inherent randomness or noise in data from the uncertainty due to models' lack of knowledge. Being able to separately quantify and report epistemic uncertainty would be highly valuable in clinical settings, as it could alert clinicians when the model is operating on unfamiliar or out-of-distribution cases.

Future studies could work on a few directions to address these limitations. First, integrating automatic tibial slope measurement models investigated in [86, 90, 119, 73] into the pipeline could help establish an end-to-end predictive framework that automates MRI feature extraction and prediction, minimizing manual annotation and enabling real-time clinical predictions. Once automated and refined, this would allow for seamless transition from small personal datasets to online ACL registries for large quantities of data, more comprehensive feature coverage, and continual learning, improving model generalization and mitigating overfitting, providing more robust identification of variables associated with ACL injury. Secondly, a comparative analysis of tibial slope measurement techniques could evaluate their robustness and reliability, potentially addressing limitations pointed out by [116]. Lastly, models capable of uncertainty quantification such as Bayesian neural networks or OOD detection mechanisms can be developed to better catch data points the model is unfamiliar with for warning clinicians as well as flagging potential improvements the prognostic system can make regarding model ignorance.

References

[1] Gregorius Airlangga and Alan Liu. A hybrid gradient boosting and neural network model for predicting urban happiness: Integrating ensemble learning with deep representation for enhanced accuracy. *Machine Learning and Knowledge Extraction*, 7(1):4, 2025.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, pages 2623–2631, Anchorage, AK, USA, 2019. Association for Computing Machinery.

[3] Abeer Al-Nafjan, Amaal Aljuhani, Arwa Alshebel, Asma Alharbi, and Atheer Alshehri. Artificial intelligence in predictive healthcare: A systematic review. *Journal of Clinical Medicine*, 14(19):6752, 2025.

[4] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 272–281. PMLR, 06 2019.

[5] Luca Andriollo, Aurelio Picchi, Rudy Sangaletti, Loris Perticarini, Stefano Marco Paolo Rossi, Giandomenico Logroscino, and Francesco Benazzo. The role of artificial intelligence in anterior cruciate ligament injuries: Current concepts and future perspectives. *Healthcare*, 12(3):300, 2024.

[6] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[7] M. J. Awan, M. T. Khan, M. Abbas, M. Khan, Z. U. Khan, and et al. Automated knee mr images segmentation of anterior cruciate ligament tear. *Sensors*, 22(4):1552, 2022.

[8] Karsyn N. Bailey, Kenneth T. Gao, Ryan T. Halvorson, Jacob F. Oeding, Sharmila Majumdar, Valentina Pedoia, and .... Association between tibiofemoral bone shape features and retears after anterior cruciate ligament reconstruction. *Orthopaedic Journal of Sports Medicine*, 12(11):23259671241289096, 2024.

[9] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France, July 2001. Association for Computational Linguistics.

[10] Steve Bayer, Sean J. Meredith, Kevin W. Wilson, Darren de Sa, Tristan Pauyo, Ken Byrne, Colum M. McDonough, and Volker Musahl. Knee morphological risk factors for anterior cruciate ligament injury: A systematic review. *The Journal of Bone & Joint Surgery. American Volume*, 102(8):703–718, 2020.

[11] D. Beynnon, Bruce W. Tourville, Thomas J. Shultz, Sandra M. Vacek, Pamela R. Sturnick, Derek Mack Gardner-Morse, A. Holterman, Laurie C. Smith, Helen R. Slauterbeck, James and J. Johnson, Robert˙ Intrinsic risk factors for first-time non-contact anterior cruciate ligament injury: A prospective study of college and high school athletes. *Journal of Sports Health*, 15(3):433–442, 2022.

[12] Katherine M. Birt and Helen E. McCarthy. The effect of estrogen on anterior cruciate ligament structure and function: A systematic review. *Journal of Musculoskeletal Disorders and Treatment*, 9(3):127, 2023.

[13] Benjamin Blackman, Prushoth Vivekanantha, Rafay Mughal, Ayoosh Pareek, Anthony Bozzo, Kristian Samuelsson, and Darren de Sa. Machine learning models predicting risk of revision or secondary knee injury after anterior cruciate ligament reconstruction demonstrate variable discriminatory and accuracy performance: a systematic review. *BMC Musculoskeletal Disorders*, 26(1):16, 2025. Epub 2025 Jan 4.

[14] Barry P. Boden, Frances T. Sheehan, Joseph S. Torg, and Timothy E. Hewett. Noncontact anterior cruciate ligament injuries: Mechanisms and risk factors. *Journal of the American Academy of Orthopaedic Surgeons*, 18(9):520–527, September 2010. PMCID: PMC3625971; PMID: 20810933.

[15] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024.

[16] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[17] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[18] Josh C. Carter, Daniel R. Sturnick, Pamela M. Vacek, Michael J. DeSarno, Erin C. Argentieri, James R. Slauterbeck, Robert J. Johnson, and Bruce D. Beynnon. Relationship between geometry of the extensor mechanism of the knee and risk of anterior cruciate ligament injury. *Journal of Orthopaedic Research*, 35(5):965–973, 2017.

[19] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016. Submitted March 9, 2016; revised June 10, 2016.

[20] Jason Corban, Justin-Pierre Lorange, Carl Laverdière, Jason Khoury, Gil Rachevsky, Mark Burman, and Paul A. Martineau. Artificial intelligence in the management of anterior cruciate ligament injuries. *Orthopaedic Journal of Sports Medicine*, 9(7):23259671211014206, 2021.

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

[22] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers: 2nd edition (with python examples). Preprint arXiv:2004.04523v2, arXiv, 2020.

[23] Boyi Dai, Dan Herman, Hua Liu, William E. Garrett, and Bing Yu. Prevention of acl injury, part i: Injury characteristics, risk factors, and loading mechanism. *Research in Sports Medicine*, 20(3-4):180–197, 2012.

[24] Rahul C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[25] Yusuke Endo, Masahiro Takemura, Masahiko Monma, Hirotaka Mutsuzaki, and Masafumi Mizukami. Difference in sex and the effect of a dominant lower extremity in the posterior tibial slope angle in healthy japanese subjects. *Asia Pacific Journal of Sports Medicine, Arthroscopy, Rehabilitation and Technology*, 23:8–12, 2020. eCollection 2021 Jan.

[26] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[27] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint*, arXiv:2506.16791, 2025. Accepted (spotlight) at NeurIPS 2025 Datasets and Benchmarks Track; v4 available.

[28] Jennifer Evans, Ahmed Mabrouk, and Jeffery L. Nielson. Anterior cruciate ligament knee injury. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island – FL, USA, jan– edition, 2023. Accessed: YYYY-MM-DD.

[29] Joshua S. Everhart, Sercan Yalçin, and Kurt P. Spindler. Twenty-year outcomes after anterior cruciate ligament reconstruction: A systematic review of prospectively collected data. *The American Journal of Sports Medicine*, 50(10):2842–2852, 2022.

[30] Evelyn Fix and L. Hodges, J.˙Discriminatory analysis: Nonparametric discrimination — consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.

[31] João Fonseca and Fernando Bação. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.

[32] Peter I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[33] A. Fältström, M. Hägglund, J. Kvist, N. F. N. Bittencourt, and L. D. Mendonça. Clinical risk profile for a second anterior cruciate ligament injury in female soccer players after anterior cruciate ligament reconstruction. *The American Journal of Sports Medicine*, 49(6):1421–1430, 2021.

[34] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[35] Mohamad Hadi Gerami, Farshid Haghi, Ferdos Pelarak, and Seyed Rohallah Mousavibaygei. Anterior cruciate ligament (acl) injuries: A review on the newest reconstruction techniques. *Journal of Family Medicine and Primary Care*, 11(3):852–856, 2022. Epub 2022 Mar 10.

[36] Amirata Ghorbani and James Y. Zou. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, page –, Vancouver, Canada, 2020. Curran Associates.

[37] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *Proceedings of the International Conference on Learning Representations (ICLR), 2025*, January 2025. Poster.

[38] Alberto Grassi, Nicola Pizza, Belal Bashar Hamdan Al-Zu'bi, Giacomo Dal Fabbro, Gian Andrea Lucidi, and Stefano Zaffagnini. Clinical outcomes and osteoarthritis at very long-term follow-up after acl reconstruction: A systematic review and meta-analysis. *Orthopaedic Journal of Sports Medicine*, 10(1):23259671211062238, 2022.

[39] C. Gregg and K. Hazelwood. Where is the data? why you cannot debate cpu vs. gpu performance without the answer. In *Proceedings of the 2011 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2011)*, pages 134–144. IEEE, 2011.

[40] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022) — Datasets and Benchmarks Track*, page —. Curran Associates, Inc., 2022.

[41] J. Hashemi, N. Chandrashekar, H. Mansouri, J. Slauterbeck, B. D. Beynnon, J. Avedschmidt, S. Rundell, W. D. McLeod, and A. Chaudhari. Shallow medial tibial plateau and steep medial and lateral tibial slopes: new risk factors for anterior cruciate ligament injuries. *The American Journal of Sports Medicine*, 38(1):54–62, 2010.

[42] Javad Hashemi, Naveen Chandrashekar, Brian Gill, Bruce D. Beynnon, James R. Slauterbeck, Jr. Schutt, Robert C., Hossein Mansouri, and Eugene Dabezies. The geometry of the tibial plateau and its influence on the biomechanics of the tibiofemoral joint. *The Journal of Bone & Joint Surgery*, 90(12):2724–2734, 2008.

[43] David Haussler and Manfred K. Warmuth. The probably approximately correct (pac) and other learning models. In *Foundations of Knowledge Acquisition*, volume 195 of *The Springer International Series in Engineering and Computer Science*, pages 291–312. Springer, New York / London, 1993.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[45] Miao He and Jie Li. Increased lateral femoral condyle ratio measured by mri is associated with higher risk of noncontact anterior cruciate ligament injury. *BMC Musculoskeletal Disorders*, 23:190, 2022.

[46] Svenja Heidecke, Nikolaus Kraml, David Haslhofer, Paul M. Schwarz, Tobias Gotterbarm, and Philipp W. Winkler. High posterior tibial slope is associated with higher failure rates in nonoperative management of primary anterior cruciate ligament injury. *Journal of Experimental Orthopaedics*, 12(4):e70502, 2025.

[47] Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018. arXiv:1806.03852 [cs.CV].

[48] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. Submitted 1 Dec 2017.

[49] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979, 2022.

[50] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[51] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[52] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[53] Shayan Hosseinzadeh, Ata M. Kiapour, et al. Sex differences in anatomic features linked to anterior cruciate ligament injuries during skeletal growth and maturation. *American Journal of Sports Medicine*, 48(9):2205–2212, 2020.

[54] Robert Hudek, Silvia Schmutz, Felix Regenfelder, Bruno Fuchs, and Peter P. Koch. Novel measurement technique of the tibial slope on conventional mri. *Clinical Orthopaedics and Related Research*, 467(8):2066–2072, 2009.

[55] Quinn J. Johnson, Mohamed S. Jabal, Alexandra M. Arguello, Yining Lu, Kevin Jurgensmeier, Bruce A. Levy, Christopher L. Camp, and Aaron J. Krych. Machine learning can accurately predict risk factors for all-cause reoperation after aclr: creating a clinical tool to improve patient counseling and outcomes. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(10):4099–4108, 2023.

[56] I. T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[57] Yusuke Kamatsuki, Marie Synnøve Qvale, Kathrin Steffen, Arnlaug Wangensteen, and Tron Krosshaug. Anatomic risk factors for initial and secondary noncontact anterior cruciate ligament injury: A prospective cohort study in 880 female elite handball and soccer players. *The American Journal of Sports Medicine*, 53(1):123–131, 2025.

[58] Axel Karlsson, Tianze Wang, Slawomir Nowaczyk, Sepideh Pashami, and Sahar Asadi. Mind the data, measuring the performance gap between tree ensembles and deep learning on tabular data. In Ioanna Miliou, Nico Piatkowski, and Panagiotis Papapetrou, editors, *Advances in Intelligent Data Analysis XXII*, volume 14641 of *Lecture Notes in Computer Science*, pages 65–76. Springer, Cham, 2024.

[59] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[60] Sefa Key, Mehmet Baygin, Sukru Demir, Sengul Dogan, and Turker Tuncer. Meniscal tear and acl injury detection model based on alexnet and iterative relieff. *Journal of Digital Imaging*, 35(–):200–212, 2022.

[61] Ata M. Kiapour, Daniel S. Yang, Gary J. Badger, Naga P. Karamchedu, Martha M. Murray, Paul D. Fadale, Michael J. Hulstyn, Robert M. Shalvoy, and Braden C. Fleming. Anatomic features of the tibial plateau predict outcomes of acl reconstruction within 7 years after surgery. *The American Journal of Sports Medicine*, 47(2):303–311, 2019.

[62] Dong-Keon Kim, DongHeum Ryu, Yongbin Lee, and Dong-Hoon Choi. Generative models for tabular data: A review. *Journal of Mechanical Science and Technology*, 38(9):4989–5005, 2024.

[63] Elife Oztürk Kiyak, Bita Ghasemkhani, and Derya Birant. High-level k-nearest neighbors (hlknn): A supervised machine learning model for classification analysis. *Electronics*, 12(18):3828, 2023. Published 10 September 2023; Received 14 August 2023; Revised 6 September 2023; Accepted 7 September 2023.

[64] Christos Kokkotis, Serafeim Moustakidis, Themistoklis Tsatalas, Charis Ntakolia, Georgios Chalatsis, Stylianos Konstadakos, Michael E. Hantes, Giannis Giakas, and Dimitrios Tsaopoulos. Leveraging explainable machine learning to identify gait biomechanical parameters associated with anterior cruciate ligament injury. *Scientific Reports*, 12:6647, 2022.

[65] Omkar P. Kulkarni, Ganesh N. Pundkar, and Satish B. Sonar. A comparative study of mri versus arthroscopic findings in acl and meniscal injuries of the knee. *International Journal of Research in Orthopaedics*, 4(2):198–202, 2018.

[66] Keelara Mahadevappa Pawan Kumar, Aniruddha Mundargi, Madhuchandra Puttamaregowda, Rakshith Kumar Kammagondanahalli, Sagar Perumalswamy, and Shetty M. Shantharam. Lfci and knee morphological parameters as a risk factor in anterior cruciate ligament tear on magnetic resonance imaging - a case control analysis. *Journal of Clinical Orthopaedics and Trauma*, page 55:102514, 2024.

[67] Christophe Ley, Kyle Martin, R. Anil Pareek, Andreas Groll, Reinhard Seil, and Thorsten Tischer. Machine learning and conventional statistics: making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(3):753–757, 2022.

[68] Kun Li, Jun Du, Li-Xin Huang, Li Ni, Tao Liu, and Hui-Lin Yang. The diagnostic accuracy of magnetic resonance imaging for anterior cruciate ligament injury in comparison to arthroscopy: a meta-analysis. *Scientific Reports*, 7(1):7583, 2017.

[69] Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *arXiv preprint arXiv:2210.05021*, 2022. v3, last revised 27 Feb 2024.

[70] Chao Liu, Xueqing Yu, Dingyu Wang, and Tingting Jiang. Aclnet: A deep learning model for acl rupture classification combined with bone morphology. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005, pages 57–67. Springer Nature Switzerland, October 2024.

[71] C. D. Lopez et al. Machine learning can accurately predict overnight stay, short-term complications, and readmission after total knee arthroplasty. *The Knee*, 38:60–69, 2023.

[72] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. PhD thesis, University of Liège, Belgium, Oct 2014. arXiv:1407.7502 [stat.ML].

[73] Y. Lu, A. Pareek, L. Yang, and et al. Deep learning artificial intelligence tool for automated radiographic determination of posterior tibial slope in patients with acl injury. *Orthopaedic Journal of Sports Medicine*, 11(12):23259671231215820, 2023.

[74] Yining Lu, Kevin Jurgensmeier, Abhinav Lamba, Linjun Yang, Mario Hevesi, Christopher L. Camp, Aaron J. Krych, and Michael J. Stuart. Posttraumatic arthritis after anterior cruciate ligament injury: Machine learning comparison between surgery and nonoperative management. *American Journal of Sports Medicine*, 53(5):1050–1060, 2025. Epub 2025 Mar 13.

[75] Yining Lu, Sara E. Till, Joshua R. Labott, Anna K. Reinholz, Mario Hevesi, Aaron J. Krych, Christopher L. Camp, and Kelechi R. Okoroha. Graft failure and contralateral acl injuries after primary acl reconstruction: An analysis of risk factors using interpretable machine learning. *Orthopaedic Journal of Sports Medicine*, 12(10):23259671241282316, 2024.

[76] Yi lun Wang, Tuo Yang, Chao Zeng, Jie Wei, Dong xing Xie, Yuan heng Yang, Hui zhong Long, Bei Xu, Yu xuan Qian, Shi de Jiang, and Guang hua Lei. Association between tibial plateau slopes and anterior cruciate ligament injury: A meta-analysis. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 33(6):1248–1259.e4, 2017.

[77] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4765–4774, 2017.

[78] Nathan A. Mall, Peter N. Chalmers, Mario Moric, Miho J. Tanaka, Brian J. Cole, Jr. Bach, Bernard R., and Jr. Paletta, George A. Incidence and trends of anterior cruciate ligament reconstruction in the united states. *The American Journal of Sports Medicine*, 42(10):2363–2370, 2014.

[79] Stephen W. Marshall. Recommendations for defining and classifying anterior cruciate ligament injuries in epidemiologic studies. *Journal of Athletic Training*, 45(5):516–518, 2010.

[80] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Benjamin Feuer, Chinmay Hegde, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? *arXiv preprint arXiv:2305.02997*, 2023. Version v4, revised 15 Jul 2024.

[81] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[82] Benjamin R. Mitchell. *The Spatial Inductive Bias of Deep Learning*. PhD thesis, Johns Hopkins University, Baltimore, Maryland, USA, March 2017. PhD Dissertation.

[83] F. R. Noyes, P. A. Mooar, D. S. Matthews, and D. L. Butler. The symptomatic anterior cruciate-deficient knee. part i: the long-term functional disability in athletically active individuals. *The Journal of Bone & Joint Surgery. American Volume*, 65(2):154–162, 1983.

[84] Bambang Parmanto, Paul W. Munro, and Howard R. Doyle. Reducing variance of committee prediction with resampling techniques. *Connection Science*, 8(3-4):405–426, 1996.

[85] K. Pasupa and W. Sunhem. A comparison between shallow and deep architecture classifiers on small dataset. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6, ISBN 978-1-5090-4139-8, 2016. IEEE.

[86] Valentina Pedoia, Drew A. Lansdown, Musa Zaid, Charles E. McCulloch, Richard Souza, C. Benjamin Ma, and Xiaojuan Li. Three-dimensional mri-based statistical shape model and application to a cohort of knees with acute acl injury. *Osteoarthritis and Cartilage*, 23(10):1695–1703, 2015.

[87] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. Software available at http://scikit-learn.sourceforge.net.

[88] Craig E. Pfeifer, Paul F. Beattie, Ryan S. Sacko, and Amy Hand. Risk factors associated with non-contact anterior cruciate ligament injury: A systematic review. *The International Journal of Sports Physical Therapy*, 13(4):575–587, 2018.

[89] M. Prabu, Sushma Chandrashekar, V. R. Duddukunta, G. Naresh, and S. Raja. The influence of tibial plateau morphometric risk factors on acl injury: A retrospective

case-control study: An indian perspective. *International Journal of Health Sciences*, 6(S2):5895–5902, 2022.

[90] Shi Qiu, Yaoting Wang, Gengyan Xing, Qiumei Pu, Zhe Zhao, and Lina Zhao. Precise and efficient measurement of tibial slope on magnetic resonance imaging (mri): two novel autonomous pipelines by traditional and deep learning algorithms. *Quantitative Imaging in Medicine and Surgery*, 14(8):5304–5320, 2024.

[91] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025) — Poster Session*, July 2025. Poster #E-3208.

[92] Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina*, 56(9):455, 2020.

[93] Ahmad Rashid, Serena Hacker, Guojun Zhang, Agustinus Kristiadi, and Pascal Poupart. Preventing arbitrarily high confidence on far-away data in point-estimated discriminative neural networks. *arXiv preprint arXiv:2311.03683*, 2023. Accepted at AISTATS 2024.

[94] Hailee Reist, Pamela M. Vacek, Nathan Endres, Timothy W. Tourville, Mathew Failla, Andrew Geeslin, Matthew Geeslin, Andy Borah, Mickey Krug, Rebecca Choquette, Mike Toth, and Bruce D. Beynnon. Risk factors for concomitant meniscal injury with sport-related anterior cruciate ligament injury. *Orthopaedic Journal of Sports Medicine*, 11(9):23259671231196492, 2023.

[95] Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, and Sebastian Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57:273, 2024. Open Access.

[96] Nathan D. Schilaty, Nathaniel A. Bates, Sydney Kruisselbrink, Aaron J. Krych, and Timothy E. Hewett. Linear discriminant analysis successfully predicts knee injury outcome from biomechanical variables. *The American Journal of Sports Medicine*, 48(10):2447–2455, 2020.

[97] A. Schulc, C.B.G. Leite, M. Csákvári, L. Lattermann, M.F. Zgoda, E.M. Farina, C. Lattermann, Z. Tősér, and G. Merkely. Identifying anterior cruciate ligament

injuries through automated video analysis of in-game motion patterns. *Orthopaedic Journal of Sports Medicine*, 12(3):23259671231221579, 2024.

[98] Torgyn Shaikhina and Natalia A. Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*, 75:51–63, 2017.

[99] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[100] Ira Shavitt and Eran Segal. Regularization learning networks: Deep learning for tabular datasets. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 1386–1396, Montréal, Canada, 2018.

[101] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *arXiv preprint*, abs/2106.03253, 2021.

[102] Helen C. Smith, Pamela Vacek, Robert J. Johnson, James R. Slauterbeck, Javad Hashemi, Sandra J. Shultz, and Bruce D. Beynnon. Risk factors for anterior cruciate ligament injury: A review of the literature—part 1: Neuromuscular and anatomic risk. *Sports Health: A Multidisciplinary Approach*, 4(1):69–78, 2012.

[103] Hanna H. Sorensen, Shreya M. Saraf, and Mary K. Mulcahey. What she's up against: The hidden challenges facing female youth athletes. *Sports Medicine Update*, Summer 2025:–, 2025. Issue: Summer 2025, American Orthopaedic Society for Sports Medicine.

[104] Daniel R. Sturnick, Pamela M. Vacek, Michael J. DeSarno, Mack G. Gardner-Morse, Timothy W. Tourville, James R. Slauterbeck, Robert J. Johnson, Sandra J. Shultz, and Bruce D. Beynnon. Combined anatomic factors predicting risk of anterior cruciate ligament injury for males and females. *The American Journal of Sports Medicine*, 43(4):839–847, 2015. PMID: 25583759.

[105] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS) 34*, pages 144–157, 2021.

[106] Zhuo Sun, Pawel Cieszczyk, and Aleksandra Bojarczuk. Col5a1 rs13946 polymorphism and anterior cruciate ligament injury: Systematic review and meta-analysis. *International Journal of Molecular Sciences*, 26(13):6340, 2025.

[107] Juri Taborri, Luca Molinaro, Adriano Santospagnuolo, Mario Vetrano, Maria Chiara Vulpiani, and Stefano Rossi. A machine-learning approach to measure the anterior cruciate ligament injury risk in female basketball players. *Sensors*, 21(9):3141, 2021.

[108] Iskandar Tamimi, Joaquín Ballesteros, Almudena Perez Lara, Jimmy Tat, Motaz Alaqueel, Justin Schupbach, Yousef Marwan, Cristina Urdiales, Jesús Manuel Gomez-de Gabriel, Mark Burman, and Paul Andre Martineau. A prediction model for primary anterior cruciate ligament injury using artificial intelligence. *Orthopaedic Journal of Sports Medicine*, 9(9):23259671211027543, 2021.

[109] V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[110] Kinjal Vasavada, Vrinda Vasavada, Jay Moran, Sai Devana, Changhee Lee, Sharon L. Hame, Laith M. Jazrawi, Orrin H. Sherman, Laura J. Huston, Amanda K. Haas, Christina R. Allen, Daniel E. Cooper, Thomas M. DeBerardino, Kurt P. Spindler, Michael J. Stuart, Annunziato Amendola, Christopher C. Annunziata, Robert A. Arciero, Bernard R. Bach, Champ L. Baker, Arthur R. Bartolozzi, Keith M. Baumgarten, Jeffrey H. Berg, Geoffrey A. Bernas, Stephen F. Brockmeier, Robert H. Brophy, Charles A. Bush-Joseph, J. Brad Butler, James L. Carey, James E. Carpenter, Brian J. Cole, Jonathan M. Cooper, Charles L. Cox, R. Alexander Creighton, Tal S. David, Warren R. Dunn, David C. Flanigan, Robert W. Frederick, Theodore J. Ganley, Charles J. Gatt, Steven R. Gecha, James Robert Giffin, Jo A. Hannafin, Norman L. Harris, Keith S. Hechtman, Elliott B. Hershman, Rudolf G. Hoellrich, David C. Johnson, Timothy S. Johnson, Morgan H. Jones, Christopher C. Kaeding, Ganesh V. Kamath, Thomas E. Klootwyk, Bruce A. Levy, C. Benjamin Ma, G. Peter Maiers, Robert G. Marx, Matthew J. Matava, Gregory M. Mathien, David R. McAllister, Eric C. McCarty, Robert G. McCormack, Bruce S. Miller, Carl W. Nissen, Daniel F. O'Neill, Brett D. Owens, Richard D. Parker, Mark L. Purnell, Arun J. Ramappa, Michael A. Rauh, Arthur C. Rettig, Jon K. Sekiya, Kevin G. Shea, James R. Slauterbeck, Matthew V. Smith, Jeffrey T. Spang, Steven J. Svoboda, Timothy N. Taft, Joachim J. Tenuta, Edwin M. Tingstad, Armando F. Vidal, Darius G. Viskontas, Richard A. White, James S. Williams, Michelle L. Wolcott, Brian R. Wolf, Rick W. Wright, James J. York, and "MARS Group". A novel machine learning model to predict revision acl reconstruction failure in the mars cohort. *Orthopaedic Journal of Sports Medicine*, 12(11):23259671241291920, November 2024.

[111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998–6008,

2017. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017).

[112] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation. *arXiv preprint arXiv:2405.09591*, May 2024. 36 pages.

[113] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. *arXiv preprint arXiv:2205.09310*, 2022. Accepted by ICML 2022.

[114] N. Wetters, A. E. Weber, T. H. Wuerz, D. L. Schub, and B. R. Mandelbaum. Mechanism of injury and risk factors for anterior cruciate ligament injury. *Operative Techniques in Sports Medicine*, 24(1):2–6, 2016.

[115] D. C. Whitney, D. R. Sturnick, P. M. Vacek, M. J. DeSarno, M. G. Gardner-Morse, T. W. Tourville, H. C. Smith, J. R. Slauterbeck, R. J. Johnson, S. J. Shultz, and B. D. Beynnon. Relationship between the risk of suffering a first-time noncontact ACL injury and geometry of the femoral notch and ACL: A prospective cohort study with a nested case-control analysis. *The American Journal of Sports Medicine*, 42(8):1796–1805, 2014. PMID: 24866891.

[116] S. C. Wordeman, C. E. Quatman, C. C. Kaeding, and T. E. Hewett. In vivo evidence for tibial plateau slope as a risk factor for anterior cruciate ligament injury: A systematic review and meta-analysis. *The American Journal of Sports Medicine*, 40(7):1673–1681, 2012.

[117] Sen Wu, Hongyang Zhang, Gregory Valiant, and Christopher Re. On the generalization effects of linear transformations in data augmentation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10410–10420. PMLR, 13–18 Jul 2020.

[118] Yang Yang, Hongjian Sun, Yu Zhang, Tiefu Zhang, Jialei Gong, Yunbo Wei, Yong-Gang Duan, Minglei Shu, Yuchen Yang, Di Wu, and Di Yu. Dimensionality reduction by umap reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Reports*, 36(4):109442, 2021.

[119] Shang-Yu Yao, Xue-Zhi Zhang, Soumyajit Podder, Chen-Te Wu, Yi-Shen Chan, Dan Berco, and Cheng-Pang Yang. Enhanced reliability and time efficiency of deep

learning - based posterior tibial slope measurement over manual techniques. *Knee Surgery, Sports Traumatology, Arthroscopy*, 33(1):59–69, 2024.

[120] Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. In *Proceedings of the 2025 International Conference on Representation Learning (ICLR)*, 2025.

[121] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at TabPFN v2: Understanding its strengths and extending its capabilities. *arXiv preprint*, arXiv:2502.17361, 2025. Submitted 24 Feb 2025; revised 11 Jun 2025.

[122] Zipeng Ye, Tianlun Zhang, Chenliang Wu, Yi Qiao, Wei Su, Jiebo Chen, Guoming Xie, Shikui Dong, Junjie Xu, and Jinzhong Zhao. Predicting the objective and subjective clinical outcomes of anterior cruciate ligament reconstruction: A machine learning analysis of 432 patients. *The American Journal of Sports Medicine*, 50(14):3786–3795, 2022.

[123] Hannah Yoo and Raghavendra Marappa-Ganeshan. *Anatomy, Bony Pelvis and Lower Limb, Knee Anterior Cruciate Ligament*. StatPearls Publishing, Treasure Island (FL), Jul 24 2023. Update January 2025.

[124] Guri Zabërgja, Arlind Kadra, Christian M. M. Frey, and Josif Grabocka. Tabular data: Is deep learning all you need? *arXiv preprint arXiv:2402.03970*, 2024. Submitted Feb. 6, 2024; v3 revised Oct. 5, 2025.

[125] Cindy Zeng, Felipe Moreira Borim, and Timothy Lording. Increased posterior tibial slope is a risk factor for anterior cruciate ligament injury and graft failure after reconstruction: A systematic review. *Journal of ISAKOS*, 12:100854, 2025.

[126] Tianlun Zhang and [add other authors as appropriate]. Ensemble algorithm for risk prediction of clinical failure after anterior cruciate ligament reconstruction. *Orthopaedic Journal of Sports Medicine*, 12:23259671241261695, 2024.

[127] Xiyuan Zhang and Danielle Maddix Robinson. Mitra: Mixed synthetic priors for enhancing tabular foundation models. https://www.amazon.science/blog/mitra-mixed-synthetic-priors-for-enhancing-tabular-foundation-models, July 2025. Amazon Science Blog.

[128] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, 2004.

APPENDICES

Appendix A

Raw Dataset

Table A.1: TAFs for ACL injury prediction [41]

Subject #	**CTS (°)**	**MTS (°)**	**LTS (°)**	**MTD (mm)**	**Sex**	**Injury**
165	4	7	12	1.47	F	Injured
176	2	4	10	1.07	F	Injured
175	4	11	10	0.83	F	Injured
169	3	14	10	0.98	F	Injured
153	0	5	10	0.77	F	Injured
168	3	6	10	0.97	F	Injured
162	2	6	9	1.05	F	Injured
180	2	6	9	0.65	F	Injured
170	4	8	8	1.31	F	Injured
177	3	10	10	1.63	F	Injured
152	2	4	5	1.33	F	Injured
171	2	14	12	1.39	F	Injured
179	5	6	9	1.9	F	Injured
161	0	4	7	1.94	F	Injured
181	0	3	7	1.7	F	Injured
173	4	7	10	2.26	F	Injured
183	0	11	10	1.25	F	Injured
186	0	4	10	2.61	F	Injured
155	1	10	9	2.27	F	Injured

Continued on next page

Table A.1: TAFs for ACL injury prediction [41]

Subject #	**CTS (°)**	**MTS (°)**	**LTS (°)**	**MTD (mm)**	**Sex**	**Injury**
185	2	4	4	1.91	F	Injured
163	0	2	8	2.19	F	Injured
154	2	10	9	2.63	F	Injured
182	4	9	10	2.5	F	Injured
158	4	9	7	2.67	F	Injured
184	0	3	9	1.71	F	Injured
160	5	9	7	2.82	F	Injured
167	2	5	7	4.22	M	Injured
156	0	6	9	0.91	M	Injured
205	1	9	9	1.01	M	Injured
204	1	10	10	1.57	M	Injured
201	1	4	8	0.83	M	Injured
202	0	1	10	1.28	M	Injured
198	3	10	9	1.93	M	Injured
203	3	9	7	2.24	M	Injured
199	5	5	10	1.37	M	Injured
188	0	10	9	0.69	M	Injured
207	6	7	9	1.97	M	Injured
208	0	4	9	1.68	M	Injured
195	6	9	9	2.08	M	Injured
200	4	5	9	2.09	M	Injured
194	0	4	7	1.97	M	Injured
191	5	2	2	3.1	M	Injured
206	7	9	9	2.84	M	Injured
193	2	8	1	2.31	M	Injured
192	1	4	8	2.45	M	Injured
187	4	6	8	3.64	M	Injured
209	0	1	8	1.83	M	Injured
157	3	0	9	3.44	M	Injured
190	2	3	2	3.82	M	Injured
4	5	10	10	3.08	F	Uninjured
29	3	8	8	1.81	F	Uninjured
11	5	7	9	2.05	F	Uninjured
8	1	9	8	1.92	F	Uninjured

Continued on next page

Table A.1: TAFs for ACL injury prediction [41]

Subject #	CTS (°)	MTS (°)	LTS (°)	MTD (mm)	Sex	Injury
7	7	10	11	2.3	F	Uninjured
27	-1	5	11	2.45	F	Uninjured
1	5	7	9	2.35	F	Uninjured
5	4	8	8	2.14	F	Uninjured
16	4	5	8	2.45	F	Uninjured
4	3	10	6	3.01	F	Uninjured
22	4	7	9	1.88	F	Uninjured
20	4	5	9	2.08	F	Uninjured
39	3	4	7	2.49	F	Uninjured
14	2	8	14	2.88	F	Uninjured
2	2	8	14	3.31	F	Uninjured
12	0	11	8	2.92	F	Uninjured
19	1	4	6	2.37	F	Uninjured
30	0	7	9	2.22	F	Uninjured
10	0	7	5	2.4	F	Uninjured
6	4	9	11	2.76	F	Uninjured
30	3	9	6	2.94	F	Uninjured
9	0	6	9	2.99	F	Uninjured
24	3	1	7	3.24	F	Uninjured
36	5	8	10	3.58	F	Uninjured
33	3	7	7	3.73	F	Uninjured
34	4	9	8	3.68	F	Uninjured
35	5	6	9	3.65	F	Uninjured
17	2	6	6	3.48	F	Uninjured
40	4	7	8	4.19	F	Uninjured
19	4	6	6	3.79	F	Uninjured
3	0	0	0	3.34	F	Uninjured
14	6	3	2	3.47	F	Uninjured
17	5	6	9	4.3	F	Uninjured
25	6	5	9	1.65	M	Uninjured
6	4	4	9	1.87	M	Uninjured
5	4	10	9	2.91	M	Uninjured
24	4	7	5	1.76	M	Uninjured
1	1	5	7	2.19	M	Uninjured

Continued on next page

Table A.1: TAFs for ACL injury prediction [41]

Subject #	**CTS (°)**	**MTS (°)**	**LTS (°)**	**MTD (mm)**	**Sex**	**Injury**
16	2	7	5	2.36	M	Uninjured
8	4	8	8	2.65	M	Uninjured
6	3	3	1	1.79	M	Uninjured
0	1	8	8	3.03	M	Uninjured
9	3	3	6	3.1	M	Uninjured
19	5	8	8	4.03	M	Uninjured
10	0	4	8	3.4	M	Uninjured
12	0	7	6	3.92	M	Uninjured
21	2	5	5	3.86	M	Uninjured
23	5	8	8	2.95	M	Uninjured
25	6	9	7	3.42	M	Uninjured
15	5	-3	5	2.55	M	Uninjured
20	3	0	0	3.67	M	Uninjured
7	5	5	1	3.75	M	Uninjured
26	0	6	6	4.07	M	Uninjured
16	3	0	4	3.82	M	Uninjured
8	6	8	4	3.23	M	Uninjured