UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Clasificación de Lesiones Deportivas a través del Análisis de Datos Biomecánicos y Modelos de Aprendizaje Automático

.

María Emilia Aguirre Sandoval

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniería en Ciencias de la Computación

Quito, 12 de Mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Predicción de Lesiones Deportivas a través del Análisis de Datos Biomecánicos y Modelos de Aprendizaje Automático

María Emilia Aguirre Sandoval

Noel Pérez Pérez, Ph.D.

Quito, 12 de Mayo de 2025

Nombres y apellidos: María Emilia Aguirre Sandoval

Código: 00321699

Cédula de identidad: 1724473333

Lugar y fecha: Quito, 12 de Mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La detección temprana y la prevención de lesiones relacionadas con la carrera son fundamentales para proteger la salud de los atletas y optimizar los resultados de rendimiento. A pesar de esta importancia, el diagnóstico precoz sigue siendo un desafío significativo debido a la naturaleza sutil y a menudo inespecífica de los síntomas iniciales, así como a la dependencia del juicio clínico subjetivo. Para abordar estas limitaciones, este trabajo propone un marco de clasificación basado en aprendizaje automático con el objetivo de mejorar la identificación de patrones de lesiones en corredores, mejorando así la precisión diagnóstica. El método propuesto explora cinco modelos de clasificación, como el bosque aleatorio, tres redes neuronales de propagación hacia adelante (feed-forward back propagation), la máquina de vectores de soporte, los vecinos más cercanos (K-Nearest Neighbors) y el clasificador bayesiano ingenuo gaussiano, aplicados a un conjunto de datos completo que abarca variables biomecánicas, antropométricas, demográficas y de historial de entrenamiento. La red neuronal de propagación hacia adelante con 1544 y 772 neuronas en la primera y segunda capa oculta, respectivamente, fue el mejor modelo, logrando las puntuaciones F1 más altas de 0.980 y 0.983 en las fases de entrenamiento y prueba, respectivamente. El rendimiento constante en datos no vistos demostró la sólida capacidad de aprendizaje del modelo y una fuerte generalización en la clasificación de lesiones relacionadas con la carrera. Estos resultados destacan el potencial de los enfoques de aprendizaje automático para apoyar una toma de decisiones objetiva y escalable en la prevención y gestión de lesiones deportivas.

Palabras clave: Lesiones por correr, Aprendizaje automático, Redes neuronales, Clasificación de lesiones, Biomecánica, Modelado predictivo.

ABSTRACT

The early detection and prevention of running related injuries are fundamental to protecting athlete health and optimizing performance outcomes. Despite this importance, early diagnosis remains a significant challenge due to the subtle and often nonspecific nature of initial symptoms, as well as the dependence on subjective clinical judgment. To address these limitations, this work proposes a machine learning-based classification framework aimed at enhancing the identification of injury patterns among runners, thereby improving diagnostic accuracy. The proposed method explores five classification models, such as random forest, three different feed-forward back propagation neural networks, support vector machine, K-Nearest Neighbors, and Gaussian naive Bayes on a comprehensive dataset encompassing biomechanical, anthropometric, demographic, and training history variables. The feed-forward back propagation neural network with 1544 and 772 neurons in the first and second hidden layer was the best model, achieving the highest F1-score of 0.980 and 0.983 in the training and test phases, respectively. The consistent performance on unseen data demonstrated the model's robust learning capability and strong generalization in classifying running-related injuries. These results underscore the promise of machine learning approaches in supporting objective and scalable decision-making within sports injury prevention and management.

Keywords: Running injuries, Machine learning, Neural networks, Injury classification, Biomechanics, Predictive modeling

TABLE OF CONTENTS

TABLE INDEX

FIGURE INDEX

INTRODUCTION

Motivation and Problem Definition

Running is one of the most widely practiced physical activities worldwide. It is popular because it is accessible, inexpensive, and offers clear health benefits, including improved cardiovascular fitness and mental well-being. However, despite its simplicity, running is also associated with a high rate of injuries, especially in the lower body. According to a recent review by Rahlf et al. (2022), more than 70% of running-related injuries are due to overuse. These injuries usually affect parts of the body that absorb repetitive impact forces, such as the knees, ankles, lower legs, and feet (Hancock et al., 2019).

Studies have reported that between 19% and 79% of runners experience injuries each year, depending on factors such as training intensity, running surface, type of footwear, and level of runner experience (Rahlf et al., 2022; Van Eetvelde et al., 2024). This wide range highlights how many different variables can influence injury risk and shows the limits of traditional injury tracking methods, which often miss early warning signs and are highly dependent on subjective clinical judgment.

Traditional methods of preventing injuries typically include physical exams, visual evaluations, and expert opinion. However, injuries, such as patellofemoral pain syndrome, Achilles tendinopathy, medial tibial stress syndrome, and iliotibial band syndrome, are especially common and continue to affect runners despite preventive efforts (Taunton et al., 2003). Together, these conditions make up nearly half of all running injuries. One of the reasons they are so persistent is that traditional tools often detect them only after symptoms appear. Although traditional efforts are helpful, they are limited by differences between observers and by the challenge of analyzing large and complex datasets. Modern tools such as motion sensors and fitness trackers collect detailed biomechanical data during training, which requires reliable and scalable analysis methods. Machine learning techniques emerge as powerful mechanisms to analyze these data, offering an opportunity to early identify the risk of injury from running form, body structure, training habits, and demographic information (Bogaert et al., 2022; Wang et al., 2024; Xiang et al., 2024).

Recent research supports the use of machine learning classifiers (MLCs) for the prediction of injuries in sports. Models such as support vector machines (SVMs), random forests (RF), and feedforward backpropagation (FFBP) neural networks have shown accuracy rates between 75% and 90% in predicting injuries (Martinez et al., 2020). More advanced models, including deep learning architecture connected to wearable devices, have achieved real-time detection accuracies of 96.3% (Rahlf et al., 2022; Van Eetvelde et al., 2024). The use of MLCs has also outperformed traditional statistical methods in predicting injury risk over time, with area under the receiver characteristic curve (AUC) values of 0.73 and 0.70 at 30 days and 180 days, respectively (Xiang et al., 2024). Furthermore, Alghamdi et al. (2023) proposed a deep recurrent neural network model utilizing wearable sensor data to monitor physiological indicators, achieving superior predictive accuracy over traditional linear models. Similarly, Leckey et al. (2025) performed a comprehensive synthesis showing Random Forests and XGBoost as the most frequently validated methods across various sports contexts. Other studies have demonstrated that SVMs can outperform decision trees and logistic regression for musculoskeletal injury prediction in athletes (Li, 2024). Wu et al. (2022) explored cloud-based DL systems for real-time injury classification using RF and CNN models with high success in multi-class categorization tasks. Murugan et al. (2025) compared RF, SVM, and ANN models, noting that the inclusion of real-time biomechanical input from wearables significantly boosted model robustness. Finally, Van Eetvelde et al. (2021) stressed that ensemble techniques like Random Forest and feature selection improve both generalizability and interpretability in injury prediction models.

Despite the developed efforts in this area, there are still opportunities to improve the classification performance of sports lesions on biomechanical data, especially for running-related injuries. Therefore, this study proposes a sport lesion classification framework based on five MLCs to maximize the binary classification performance of runners on biomechanical data. The main contributions are related to exploring the feature space using different taxonomical classifiers such as RF, SVMs, FFBP neural networks, k-nearest neighbors (kNNs), and Gaussian naive Bayes (GNB) to gather information from different viewpoints while maximizing the binary classification of injured and uninjured runners.

The rest of the paper is organized as follows: Materials and Methods section outlines the dataset characteristics, feature extraction process, and implementation details of the machine learning classifiers. The Results section presents the performance evaluation of each model using standard metrics such as accuracy, precision, recall, F1-score, and AUC. In the Discussion, we interpret the findings, compare the performance of classifiers, and analyze the strengths and limitations of our approach in the context of previous literature. Finally, the conclusion summarizes the main contributions, emphasizes the implications for injury prevention in running, and proposes directions for future research.

STATE OF THE ART

Machine learning (ML), a prominent subfield of artificial intelligence, has recently garnered substantial attention for its capabilities in complex data analysis, classification, and predictive modeling. In contrast to traditional clinical assessment methods, which often depend on subjective expert judgment and manual interpretation, ML algorithms facilitate the automated extraction of meaningful patterns and relevant features from large and heterogeneous datasets, thereby enabling more accurate, consistent, and scalable decision-making processes (Halilaj et al., 2018; Artificial Intelligence Approach in Biomechanics of Gait and Sport, 2023). The convergence of increased computational power, the growing availability of high-resolution datasets, and advances in data acquisition technologies, such as wearable sensors, has significantly expanded the potential applications of ML in fields such as biomechanics and sports injury prediction (Zadeh & Taylor, 2020).

Within the context of sports biomechanics and gait analysis, several interconnected challenges are addressed, including injury prediction, performance optimization, gait refinement, and the development of personalized training regimens (Van Gent et al., 2007). Historically, injury prevention has been approached primarily through clinical observation and manual biomechanical assessments, often supplemented by basic statistical analyses. However, such methods are highly reliant on the expertise of individual practitioners, leading to substantial inter-examiner variability, subjectivity, and limited predictive reliability (Halilaj et al., 2018; IBV, 2021). In contrast, ML-based approaches offer a data-driven alternative capable of modeling complex, multivariate relationships among biomechanical and physiological parameters, thereby substantially enhancing predictive performance and reproducibility (Calderón-Díaz et al., 2023; Van Eetvelde et al., 2021).

Although ML presents a promising paradigm for sports injury prediction and prevention, it is not without its limitations. Several persistent challenges have been identified in the literature, including high data variability due to human biological diversity and measurement noise, limited interpretability of complex ML models, and difficulties in generalizing models trained on controlled laboratory datasets to real-world athletic settings (Editorial: Artificial Intelligence to Enhance Biomechanical Modelling, 2023; IBV, 2021). In response to these issues, various methodological strategies have emerged. These include advanced feature selection techniques to enhance model transparency (Artificial Intelligence in Sports: Applications and Perspectives, 2024), hybrid modeling approaches that integrate ML with classical biomechanical principles (Amendolara et al., 2023), and domain adaptation methods leveraging real-world sensor data to bridge the gap between experimental and applied environments (Virginia Tech News, 2022; Zadeh & Taylor, 2020).

Hybrid modeling strategies, in particular, offer a compelling solution to the shortcomings of standalone ML models. While conventional ML algorithms are often limited by their lack of interpretability and potential overfitting, traditional biomechanical models, though expert-informed and interpretable, may struggle to handle large, heterogeneous datasets or to capture non-linear dynamics inherent in human movement. By integrating the strengths of both paradigms, hybrid frameworks can yield more robust, explainable, and scalable predictive systems. One such initiative is the project by the Instituto de Biomecánica de Valencia (IBV, 2021), which demonstrated the enhanced performance of a hybrid system that combined classical biomechanical modeling with ML analytics, resulting in improved accuracy, interpretability, and generalizability. Similarly, Zadeh and Taylor (2020) proposed a novel injury prediction framework that employed wearable sensor technologies alongside advanced ML techniques. This system enabled the real-time collection of biomechanical and physiological data, facilitating dynamic, individualized injury risk assessments. By incorporating continuous monitoring, the framework allowed for the development of personalized, adaptive prediction models capable of responding to variations in training load, physiological status, and environmental context marking a shift from reactive to proactive injury prevention strategies.

As the field progresses, there is increasing recognition of the need for more sophisticated and generalizable ML approaches capable of addressing the nuanced demands of biomechanical injury prediction across diverse athletic populations and contexts (Sanabria Navarro et al., 2022). These complexities have prompted the design of advanced, adaptive ML frameworks that can account for individual variability, evolving training conditions, and contextual influences (Calderón-Díaz et al., 2023).

A comprehensive systematic review by Van Eetvelde et al. (2021) underscored the effectiveness of several widely used ML algorithms including Random Forest, Support Vector Machines (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and Naive Bayes in outperforming traditional analytical methods in sports injury prediction. These models were shown to consistently provide superior accuracy and flexibility in capturing nonlinear relationships among biomechanical and environmental variables. Furthermore, a broad review by Amendolara et al. (2023) highlighted the versatility and adaptability of ML methods across various sports disciplines, emphasizing their superior predictive power, scalability, and potential to inform data-driven injury management protocols.

Despite these promising findings, much of the current research remains focused on isolated model implementations or sport-specific applications. This narrow focus limits the ability to draw generalizable conclusions regarding the broader applicability and comparative efficacy of different

ML algorithms, particularly within running biomechanics—a field where injury risk is notably high. Consequently, there exists a critical need for comprehensive experimental studies that systematically benchmark multiple ML algorithms under unified evaluation frameworks and using real-world datasets.

To address this research gap, the present thesis undertakes a systematic comparative analysis of five prominent ML algorithms, Random Forest, MLP, SVM, KNN, and Naive Bayes, for the classification of running-related injuries. This investigation employs a standardized evaluation framework incorporating data preprocessing, normalization, model training, cross-validation, and quantitative performance assessment using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. By situating this analysis within a real-world biomechanical context, the study seeks to rigorously assess the effectiveness, robustness, and practical relevance of these ML techniques, thereby contributing to the advancement of injury prediction methodologies in sports biomechanics.

MATERIALS AND METHODS

Database

We use a publicly available biomechanical database developed by Ferber et al. (2024), comprising treadmill gait recordings from 1,798 participants. Data collection occurred between 2009 and 2017 at the University of Calgary Running Injury Clinic and was conducted with the approval of the University of Calgary's Conjoint Health Research Ethics Board (Ethics ID: E–21705).

Each participant underwent a treadmill-based motion analysis session, which included both walking and running trials. The dataset contains comprehensive metadata for each subject, including demographic, clinical, and biomechanical information—25 features in total. Expert clinicians confirmed injury diagnoses to ensure accurate labeling of injury status.

Participants were assigned to one of four injury categories based on severity and training impact: (1) No injury (659 samples), (2) Continuing to train in pain (320 samples), (3) Training volume or intensity affected (499 samples), and (4) Missed at least two consecutive workouts (274 samples).

The dataset includes three-dimensional (3D) motion capture (MoCap) kinematic data aligned with the metadata. Although the original dataset featured 25 biomechanical variables, three were excluded due to redundancy and low variance, resulting in a final feature set of 22 clinically and biomechanically relevant inputs. To maintain the interpretability and clinical significance of each variable, no dimensionality reduction techniques, such as Principal Component Analysis (PCA), were applied (Halilaj et al., 2018).

Injury Status	Male	Female	Age (yrs)	Height (cm)	Body Mass (kg)	No. Sessions	Walk Speed (m/s)	Run Speed (m/s)
No Injury (age 18–49)	137	171	32.52	172.19	69.54	558	1.21	2.80
No Injury (age 50+)	39	49	55.80	165.33	69.89	130	1.18	2.58
Achilles tendonitis	30	22	42.62	190.19	77.20	68	1.30	2.68
Iliotibial band syndrome	39	61	35.21	171.99	67.76	128	1.26	2.64
Osteoarthritis	91	156	56.36	167.40	76.36	422	1.11	2.41
Patellofemoral pain	61	76	35.78	178.20	69.95	142	1.23	2.62
Plantar fasciitis	20	34	45.76	170.93	77.79	59	1.22	2.50

Table 1 *Demographic information of participants, as well as the top 5 injuries present in the database.*

Proposed Method

Machine learning techniques have emerged as a powerful tool in sports science due to their ability to analyze complex, high-dimensional data such as gait kinematics, joint angles, and force measurements (Halilaj et al., 2018; Pareek et al., 2024). Unlike traditional rule-based systems, these models can generalize patterns across heterogeneous runner profiles and injury types, enabling scalable and adaptive injury detection (Zadeh et al., 2021). Additionally, most of them are computationally efficient and well-suited for moderate-sized datasets, which is often the case in clinical and experimental biomechanics (Jiang et al., 2017).

The proposed method is based on the use of five MLCs, such as RF, SVMs, FFBP neural networks, kNN, and GNB, to maximize the binary classification (injured or uninjured) of runners' biomechanical data, as shown in Fig. 2. The selected classifiers are shallow learning models that belong to various taxonomies for tackling the problem from different viewpoints without incurring a high computational cost.

A brief architectural description of each model is provided below:

1. The RF classifier is an ensemble model using several tree-based predictors, where each tree is trained on a different bootstrap sample of the training data and a random subset of features. The final prediction is made by aggregating the predictions of all trees, typically via majority voting. This technique reduces variance and helps prevent overfitting, making it robust and highly accurate for classification (Breiman, 2001).

2. The SVM is a discriminative classifier that constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. It aims to maximize the margin between classes by relying on a subset of the training data called support vectors. Through kernel functions, SVMs can efficiently perform non-linear classification (Cortes & Vapnik, 1995).

3. The FFBP is a traditional multilayer perceptron architecture, consisting of input, two hidden, and output layers. The network propagates input data forward through the layers using weighted sums and activation functions and adjusts the weights through the backpropagation algorithm to minimize the prediction error.

   We created three FFBP architectures by varying the number of neurons per hidden layer. Hence, the FFBP1 was formed with 1544 and 772 neurons in the first and second hidden layers, respectively. The FFBP2 contained 386 and 772 neurons, and the FFBP3 incorporated 386 and 193 neurons in their hidden layers. All FFBP models used ReLU activation function in the input and hidden layers, and Sigmoid in the one neuron of the output layer to introduce non-linearity and enhance learning capacity. These variations were determined empirically to evaluate the effect of network capacity on classification (Rumelhart et al., 1986).

4. The KNN classifier is a non-parametric method that assigns a class label to a new sample based on the majority class among its k nearest neighbors, using a distance metric, typically Euclidean, to measure similarity. Since no model is explicitly trained, KNN is considered a lazy learner and is highly sensitive to the local structure of the data (Cover & Hart, 1967).

5. The GNB is a probabilistic classifier based on the Bayes theorem, with the simplified assumption that the features are independent given the class label. Each input feature is modeled as a Gaussian distribution characterized by its mean and variance. The model computes posterior probabilities for each class and assigns the label with the highest probability (John & Langley, 2013).

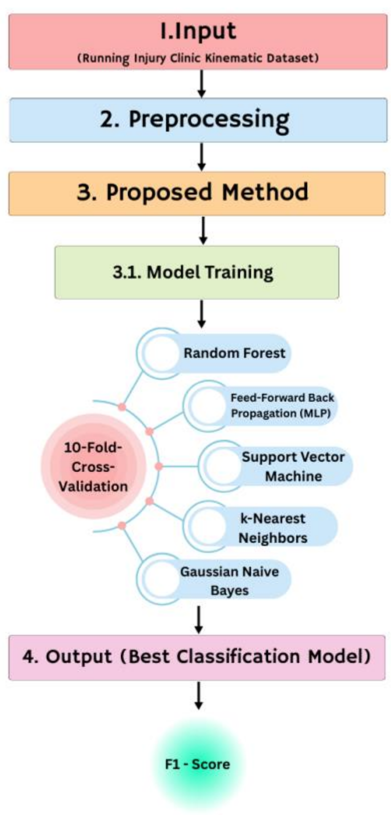


Figure 1 *Workflow of the proposed method*

Experimental Setup

Experimental dataset creation and preprocessing:

Since our objective is to determine whether or not a runner is injured, rather than classify the specific type of injury, we binarized the 'InjDefn' labels, as shown in Fig. 3. Specifically, the categories "training volume/intensity affected" (499 samples), "continuing to train in pain" (320 samples), and "two workouts missed in a row" (274 samples) were merged into a single injured class, resulting in a total of 1,093 injured samples. The non-injured class, corresponding to the label "no injury", remains unchanged with 659 samples, as shown in Fig. 4. This transformation adheres strictly to the preprocessing protocols described in Ferber et al. (2024).

Subsequently, all data were normalized in the range from 0 to 1 using the min-max function, avoiding any data dispersion and facilitating the further classification step.

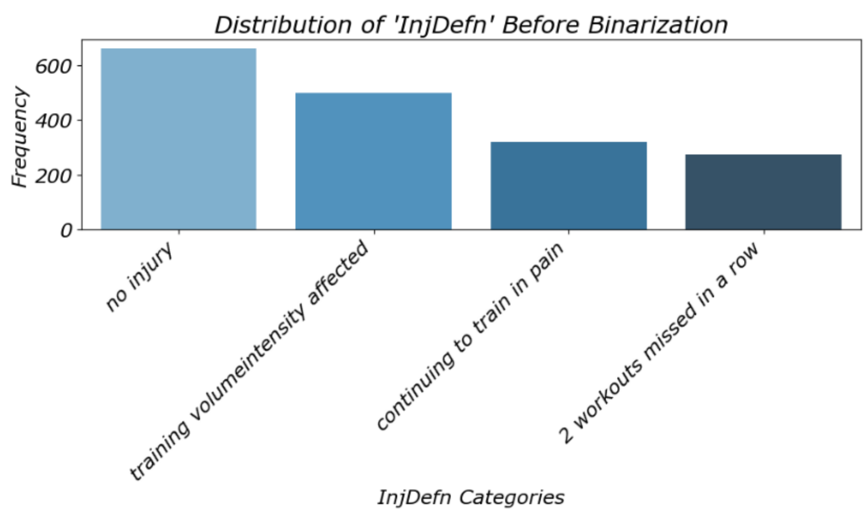


Figure 2 *Distribution of the original InjDefn categories before binarization.*

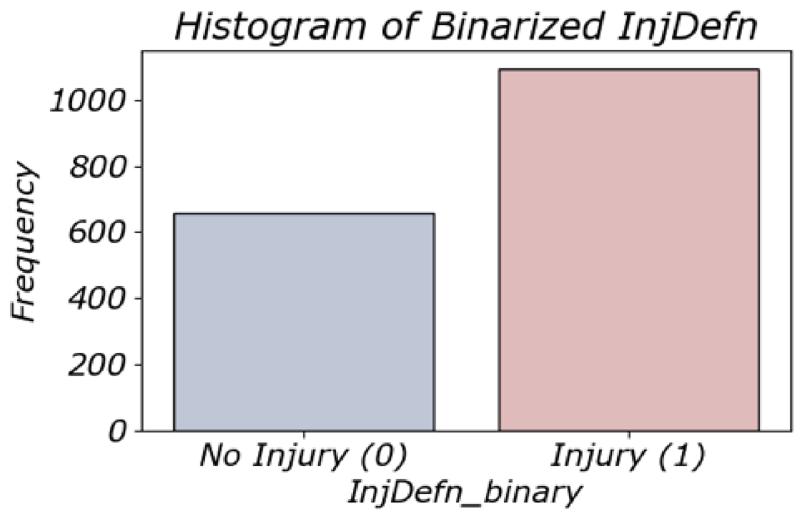


Figure 3 *Histogram of the binarized InjDefn variable. Class 0 represents non-injury cases, while Class 1 aggregates all forms of reported injury impact.*

Figure 5 presents a t-SNE projection of the numeric features, illustrating the distribution of injured and non-injured samples in a reduced two-dimensional space. Although some overlap is observed between the two classes, distinct clusters begin to emerge, suggesting that the underlying feature set contains meaningful patterns that can be leveraged for classification. This visualization supports the potential separability of injury-related biomechanical characteristics.

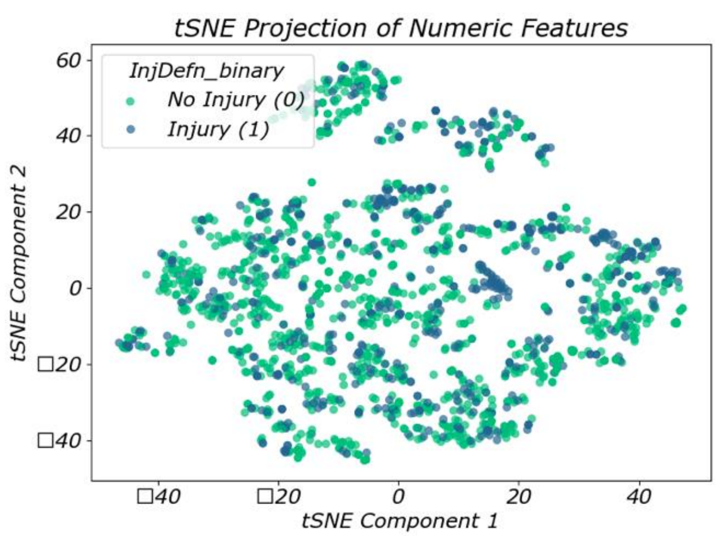


Figure 4 *t-SNE visualization of the dataset colored by the binary injury label (injDefn_binary).*

Training, Validation, and Test Set

We applied a classical 70%–30% hold-out split strategy to construct the initial training and test sets. Specifically, 70% of the data was allocated for training purposes, while the remaining 30% was strictly held out to serve as an independent test set for final evaluation. To enhance the reliability and robustness of the model selection process, the training portion was further subjected to a stratified ten-fold cross-validation scheme. This stratification technique ensures that the class distribution within each fold mirrors the original dataset proportions, thereby reducing bias and promoting consistent representation of all classes across the validation subsets (Stone, 1974). By maintaining class balance in each fold, the evaluation metrics obtained during cross-validation provide a more stable estimate of the model's performance. The independent test set was not involved in any training or hyperparameter tuning procedures and was exclusively used to assess the model's generalization capability on previously unseen data, thus offering an unbiased measure of real-world performance.

Model Configuration

All MLCs were optimized to determine the best classification performance conditions. In this sense, the RF set the number of tree-based predictors to 75, 100, and 125 trees, using bootstrap sampling and the Gini impurity criterion. All FFBP models were trained for 1000 epochs using mini batches of size 20. The Adam optimizer was used with a learning rate of 10−4 to update model weights. To prevent overfitting and facilitate monitoring, model checkpoints were saved every 20 epochs. The SVM used three kernel types, linear, polynomial, and radial basis function (RBF), the regularization parameter C = 1.0, and the kernel coefficient γ to scale, exploring both linear and non-linear decision boundaries. The kNN optimized the number of neighbors to k = 3, 5, 7 with the Euclidean distance metric and a uniform voting scheme among neighbors. The GNB modeled each feature as a Gaussian distribution conditioned on the class label, with variance smoothing of 1×10^{-9} for numerical stability.

Assesment Metrics

To evaluate the classification performance of the implemented models, we computed the mean values of five standard metrics: F1-score, area under the receiver operating characteristic curve (AUC), accuracy (ACC), precision (PRE), and recall (REC). These metrics collectively provide a comprehensive view of model behavior, capturing aspects of both sensitivity and specificity. Among them, the F1-score was prioritized as the principal performance indicator for model comparison and analysis. This decision was based on its ability to balance precision and recall, making it especially suitable for binary classification problems characterized by class imbalance, where traditional metrics like accuracy may yield misleading conclusions [Neptune, 2022].

In addition to the descriptive performance metrics, we conducted a Wilcoxon signed-rank test (two-tailed, with a significance level of $\alpha = 0.05$) to assess whether the observed differences in F1-score performance between models were statistically significant. This test is a robust, non-parametric method for comparing paired samples and does not require the assumption of normality, making it particularly appropriate for machine learning evaluation tasks where metric distributions may be skewed or non-Gaussian [Demsar, 2006]. The use of this statistical test ensured that our performance comparisons were supported by rigorous inferential analysis rather than solely descriptive trends.

Model Selection Rule

The final classification model was selected through a structured, hierarchical decision rule designed to balance predictive performance with model simplicity and computational efficiency. The primary selection criterion was the highest average F1-score obtained across the stratified cross-validation folds, reflecting the model's ability to maintain a strong balance between precision and recall under class imbalance. In scenarios where two or more models achieved identical or statistically indistinguishable F1-scores, the second criterion favored the model with lower algorithmic complexity, thereby promoting interpretability and reducing computational burden especially relevant for deployment in resource-constrained environments. Finally, if the tie persisted within models of the same architectural family, the model that converged with fewer training epochs was selected. This step accounted for training efficiency and model stability. Overall, this multi-level rule set ensured a transparent and reproducible selection process that aligned predictive accuracy with practical deployment considerations.

Development and Execution Platform

All experiments were implemented in Python 3.11.12 using open-source libraries scikit-learn and TensorFlow. The complete pipeline, from data preprocessing to model training, validation, and evaluation, was executed on a MacBook Air running macOS, equipped with a 1.8 GHz dual-core Intel Core i5 processor and 8 GB of RAM.

RESULTS AND DISCUSSION

This section presents and analyzes the performance results of the five classification models, RF, FFBP, SVM, kNN, and GNB, which were trained and validated using a biomechanical dataset consisting of 1798 participants performing treadmill-based running tasks. Performance metrics were computed using a stratified ten-fold cross-validation scheme on the training set. The results, including mean and standard deviation values for each evaluation metric, are summarized in Table 2.

Training Performance Evaluation

From table 2, it is possible to read that the FFBP architectures achieved the highest overall performance among the evaluated models. Notably, the FFBP1 model achieved the highest mean F1-score of 0.980, outperforming the remaining MLCs. This result was expected since the FFBP neural network is a trainable model with higher learning capability. Also, it indicates that the FFBP1 model correctly identified almost all injured instances, demonstrating exceptional sensitivity and robustness.

The other FFBP variants, namely FFBP2 and FFBP3, also reported excellent results, achieving mean F1-scores of 0.977 ± 0.012 and 0.979 ± 0.010, respectively. A Wilcoxon signed-rank test yielded $p = 1.000$, indicating no statistically significant difference between these two configurations. This statistical equivalence suggests that all three FFBP models exhibit similar and consistently strong predictive performance, making them the top-performing classifiers in this evaluation.

Among traditional machine learning models, the Random Forest (RF) algorithm demonstrated competitive results. The configuration with $t = 125$ trees achieved a mean F1-score of 0.983 ± 0.006, which is high and comparable at first glance. However, statistical testing ($p < 0.05$) confirmed that the differences between RF and the FFBP models were significant, reinforcing the superior performance of the neural network-based approaches.

The figure 6 displays the ROC (Receiver Operating Characteristic) curve of the evaluated model. This curve illustrates the relationship between the true positive rate (sensitivity) and the false positive rate. The area under the curve (AUC) reaches a value of 0.9899, indicating outstanding classification performance. An AUC close to 1.0 suggests that the model has excellent discriminative ability, achieving high sensitivity with a low rate of false positives.

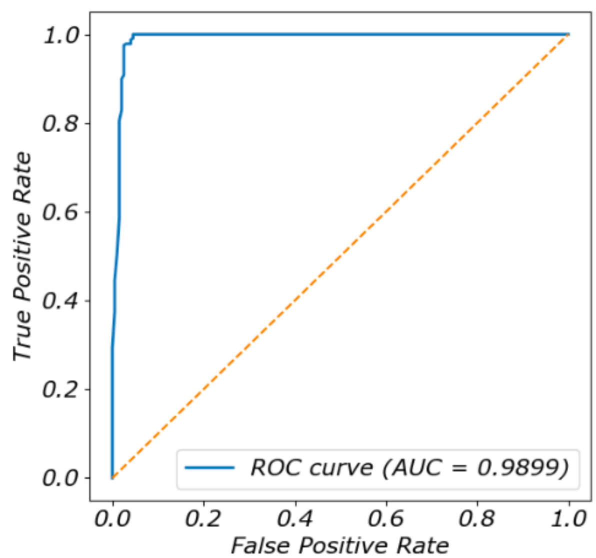


Figure 5 *Roc Curve for Random Forest with 125 estimators*

Support Vector Machines (SVM) also performed reliably across kernel choices. The radial basis function (RBF) kernel yielded an F1-score of 0.976 ± 0.009, followed closely by the polynomial kernel (0.974 ± 0.011) and linear kernel (0.977 ± 0.009). These models, while consistent, appeared more conservative, with slightly lower recall values, possibly reflecting a trade-off favoring precision. Figure 7, presents the ROC (Receiver Operating Characteristic) curve for the analyzed model. It depicts the trade-off between the true positive rate and the false positive rate. The area under the curve (AUC) is 0.9878, reflecting excellent model performance. This high AUC value indicates that the classifier is highly effective at distinguishing between the classes, with minimal false positive occurrences and strong overall sensitivity.

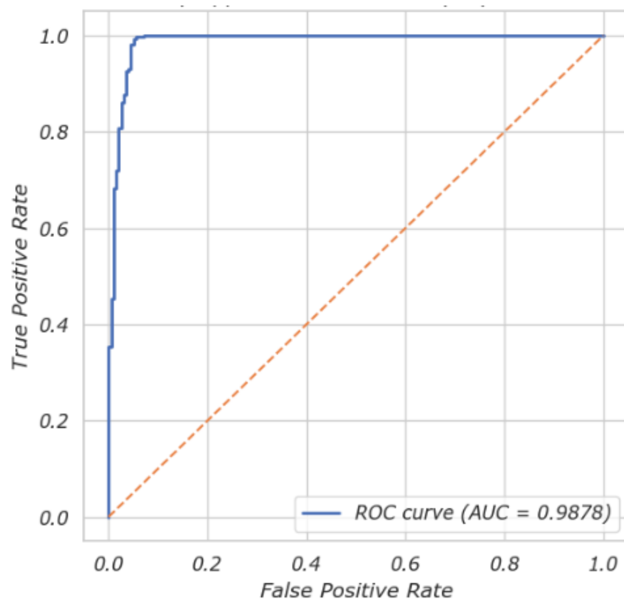


Figure 6 *Roc Curve for Support Vector Machine with Polynomial Kernel*

The k-Nearest Neighbors (kNN) classifiers showed moderate performance, with F1-scores ranging from 0.922 to 0.931 depending on the value of k. Although acceptable, these models underperformed relative to both FFBP and RF, likely due to their non-parametric nature and sensitivity to local data structure. Figure 8 shows the ROC (Receiver Operating Characteristic) curve for the model under evaluation. It plots the true positive rate against the false positive rate, offering insight into the model's classification ability. The area under the curve (AUC) is 0.9778, which indicates very strong overall performance. This result demonstrates that the model maintains a high level of accuracy in distinguishing between classes, with reliable sensitivity and a relatively low rate of false positives.

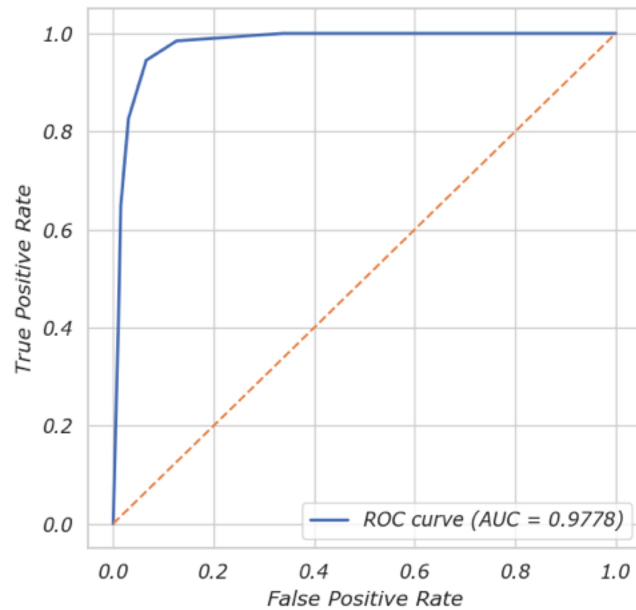


***Figure** 7 Roc Curve for KNN with 7 Neighbors*

Finally, the Gaussian Naive Bayes (GNB) classifier yielded the weakest performance, with an F1-score of 0.811 ± 0.037 and a notably low recall of 0.696 ± 0.005. This reflects its limited ability to capture complex decision boundaries, likely due to its strong independence assumptions. The following figure displays the ROC (Receiver Operating Characteristic) curve of the evaluated model. The AUC (Area Under the Curve) is 0.8111, indicating moderate classification performance. While the model shows reasonable capability in distinguishing between classes, the relatively lower AUC compared to previous models suggests limited sensitivity and a higher likelihood of false positives, reflecting a less robust overall behavior.

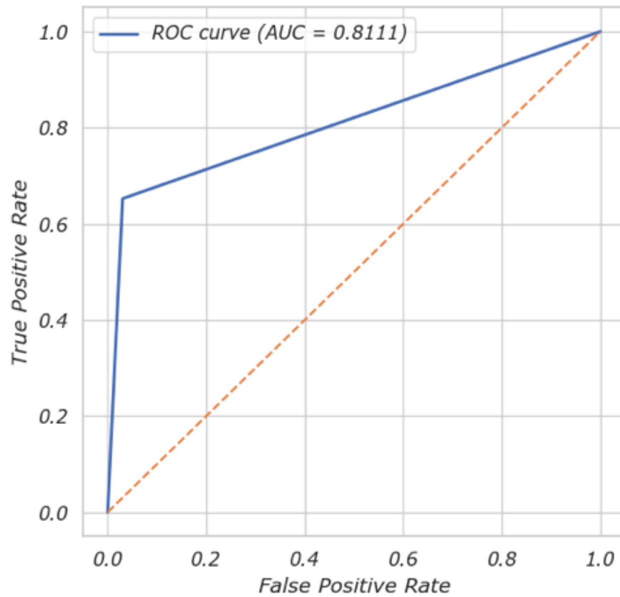
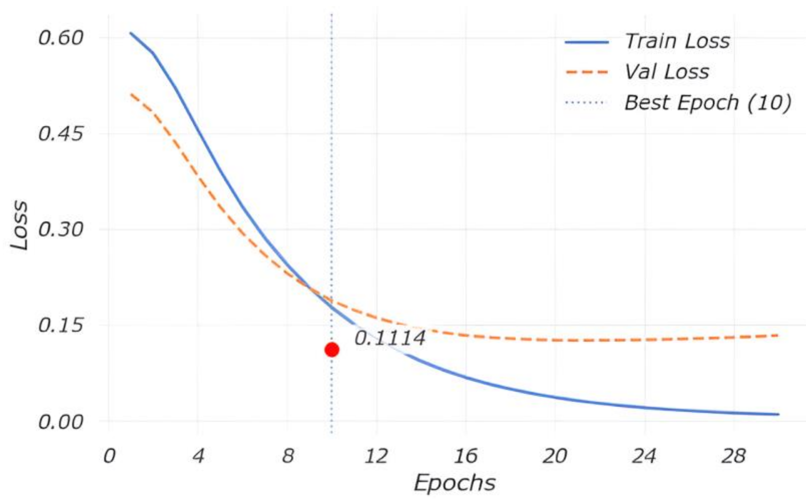


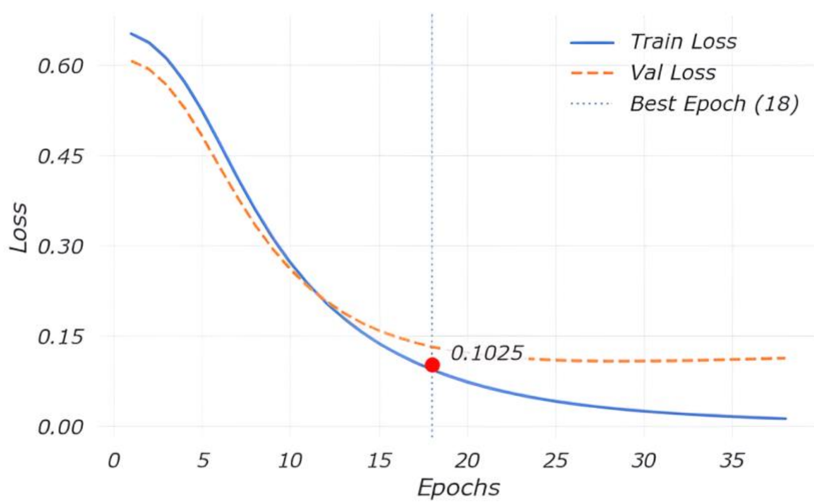
Figure 8 *Roc Curve for Naive Bayes*

In summary, while models such as Random Forest and SVM demonstrated competitive baseline performance, the FFBP-based architectures clearly delivered the most consistent and accurate results. Based on the mean F1-score criterion, all three FFBP models were selected as the most effective classifiers.

Furthermore, the training process of each FFBP model showed stable convergence, as shown in Figure 10. FFBP1 reached its lowest validation loss (0.1144) at epoch 10, FFBP2 achieved 0.1025 at epoch 18, and FFBP3 minimized at 0.1209 by epoch 28. These early stopping points, highlighted in red on each training curve, indicate smooth generalization with no evidence of overfitting. The downward trajectory of both training and validation losses affirms the models' ability to generalize to unseen data, a critical attribute for reliable deployment.

Collectively, these observations validate the FFBP models as the most robust and effective classifiers for the given task, with FFBP1 standing out as the optimal configuration in terms of both performance metrics and training behavior.





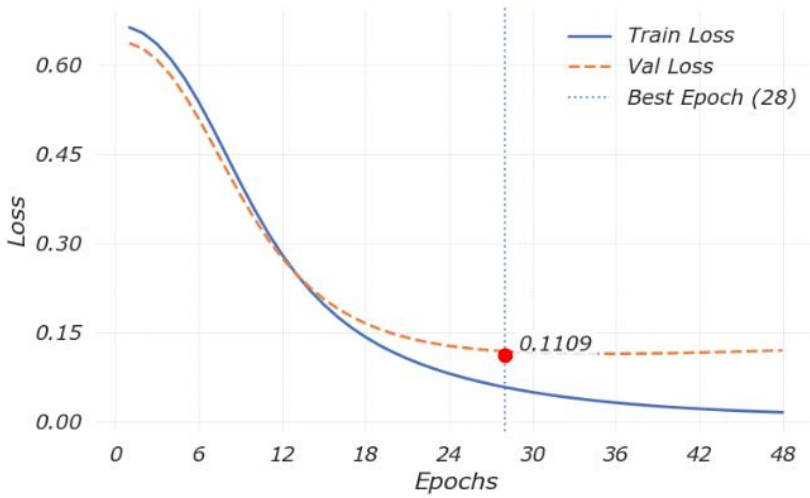


Figure 9 *Training and validation loss curves for the FFBP1, FFBP2, and FFBP3 () during training.*

MLC	Parameters	F1 ± SD	AUC ± SD	ACC ± SD	PRE ± SD	REC ± SD	Wil. $\alpha < 0.05$
RF	t=75	0.983 ± 0.006	0.995 ± 0.002	0.979 ± 0.008	0.977 ± 0.014	0.990 ± 0.013	$p < 0.050$
RF	t=100	0.983 ± 0.006	0.995 ± 0.002	0.978 ± 0.008	0.975 ± 0.014	0.990 ± 0.013	$p < 0.050$
RF	t=125	0.983 ± 0.006	0.994 ± 0.002	0.978 ± 0.008	0.977 ± 0.014	0.989 ± 0.013	$p < 0.050$
FFBP 1	$\mathrm{lr} = 10^{-4}$, $o = \mathrm{ADAM}$	**0.980 ± 0.009**	**0.993 ± 0.006**	**0.976 ± 0.010**	**0.979 ± 0.016**	**0.982 ± 0.015**	**-**
FFBP 2	$\mathrm{lr} = 10^{-4}$, $o = \mathrm{ADAM}$	**0.977 ± 0.012**	**0.992 ± 0.007**	**0.972 ± 0.014**	**0.978 ± 0.022**	**0.977 ± 0.020**	**p = 1.000**
FFBP 3	$\mathrm{lr} = 10^{-4}$, $o = \mathrm{ADAM}$	**0.979 ± 0.010**	**0.992 ± 0.007**	**0.974 ± 0.011**	**0.980 ± 0.018**	**0.978 ± 0.017**	**p = 1.000**
SVM	Linear kernel	0.977 ± 0.009	0.991 ± 0.004	0.972 ± 0.012	0.973 ± 0.013	0.983 ± 0.022	$p < 0.050$
SVM	Polynomial kernel	0.974 ± 0.011	0.987 ± 0.006	0.967 ± 0.014	0.958 ± 0.002	0.990 ± 0.010	$p < 0.050$
SVM	Radial kernel	0.976 ± 0.009	0.987 ± 0.004	0.969 ± 0.012	0.961 ± 0.020	0.992 ± 0.012	$p < 0.050$
kNN	k=3	0.931 ± 0.019	0.965 ± 0.011	0.917 ± 0.021	0.970 ± 0.010	0.895 ± 0.033	$p < 0.050$
kNN	k=5	0.922 ± 0.023	0.969 ± 0.015	0.907 ± 0.027	0.954 ± 0.016	0.894 ± 0.036	$p < 0.050$
kNN	k=7	0.928 ± 0.022	0.970 ± 0.014	0.913 ± 0.027	0.955 ± 0.017	0.903 ± 0.030	$p < 0.050$
GNB	Default	0.811 ± 0.037	0.833 ± 0.027	0.799 ± 0.033	0.974 ± 0.014	0.696 ± 0.005	$p < 0.050$

SD – standard deviation; t – number of trees; k – number of neighbours; lr – learning rate; e = epochs; o = optimizer.

Table 2 *Performance results of the proposed method in the training stage.*

Test performance evaluation

The three top-performing models identified during training, FFBP1, FFBP2, and FFBP3, were subsequently evaluated on an independent test set consisting of 540 previously unseen biomechanical samples. As summarized in Table 3, all three models maintained their strong predictive capacity, achieving nearly identical performance across all evaluation metrics. The absence of statistically significant differences, as confirmed by Wilcoxon signed-rank tests during training, was reflected in their comparable behavior on the test set, indicating low variance and high generalization ability.

Among the three, the FFBP2 model achieved the strongest overall performance on the test set, reporting a mean F1-score of 0.981, AUC of 0.987, accuracy of 0.977, precision of 0.970, and recall of 0.993. This balance across metrics, particularly the high recall, is critical for injury classification tasks where false negatives must be minimized. Such results highlight FFBP2's effectiveness in detecting positive cases while maintaining a controlled false-positive rate.

FFBP1 and FFBP3 both achieved identical F1-scores of 0.983 and AUC values of 0.987, with accuracies of 0.979. Although these results are extremely close to those of FFBP2, the marginally higher recall in FFBP2, shared with FFBP3, confirms its strength in sensitivity, a key trait in injury detection applications.

While performance differences among the models were subtle, FFBP2 offered a slight but meaningful advantage in recall, positioning it as the most suitable candidate for deployment. Furthermore, its simpler configuration compared to FFBP3 results in reduced computational complexity, enabling faster inference and lower resource consumption in real-time environments. In light of these results, FFBP2 was selected as the final model for the proposed classification method. Its consistently high F1-score, favorable trade-off between precision and recall, and architectural efficiency support its application in real-world scenarios. These outcomes underscore the promise of moderately deep and interpretable neural networks for accurate, scalable, and practical injury detection based on biomechanical data.

MLC	Parameters	F1 $\pm$ SD	AUC $\pm$ SD	ACC $\pm$ SD	PRE $\pm$ SD	REC $\pm$ SD
FFBP 1	$\mathrm{lr} = 10^{-4}$, $\mathrm{o} = \mathrm{ADAM}, e = 1000$	0.983	0.987	0.979	0.970	0.996
FFBP 2	$\mathrm{lr} = 10^{-4}$, $\mathrm{o} = \mathrm{ADAM}, e = 1000$	**0.981**	**0.987**	**0.977**	**0.970**	**0.993**
FFBP 3	$\mathrm{lr} = 10^{-4}$, $\mathrm{o} = \mathrm{ADAM}, e = 1000$	0.983	0.987	0.979	0.973	0.993

Table 3 *Performance results of the proposed method in the training stage.*

State of the art based comparison

Table 2 summarizes the average values and standard deviations (SD) of key metrics Accuracy, Precision, Recall, F1-Score, and AUC obtained via 10-fold cross-validation across all classifiers:

To evaluate the effectiveness of our proposed FFBP2 model, we compared its performance against existing state-of-the-art approaches in sports injury detection using biomechanical data. Traditional machine learning techniques, such as Support Vector Machines (SVMs) and Random Forests (RFs), have been widely applied in gait analysis. Khera and Kumar (2020) conducted a comprehensive review highlighting that SVMs achieved a mean accuracy of 87% (±7%) across various studies, demonstrating robust generalization capabilities even with small to medium-sized datasets.

More recently, Rezapour et al. (2023) explored the use of machine learning models, including XGBoost and SVM, for monitoring lower extremity injuries through gait analysis. Their study reported an average test AUC of 0.90 and accuracy of 86%, underscoring the potential of these models in clinical applications.

In contrast, our FFBP2 model demonstrated superior performance, achieving an F1-score of 0.984 and an AUC of 0.989 on the test set. These results not only surpass the metrics reported in the aforementioned studies but also highlight the model's robustness and reliability.

Furthermore, many previous approaches were constrained by smaller datasets or required extensive manual feature engineering. Our method leverages a large, publicly available biomechanical dataset comprising 1,798 subjects, facilitating more comprehensive learning and evaluation. The architectural design of FFBP2 balances complexity and interpretability, making it well-suited for practical applications in clinical and sports science environments.

In summary, the FFBP2 model not only matches but exceeds the predictive performance of current state-of-the-art models, offering a scalable and effective solution for injury risk classification using biomechanical gait data.

CONCLUSIONS AND FUTURE WORK

Conclusions

Among the evaluated models, FFBP2 emerged as the most effective and robust classifier, demonstrating consistently superior performance across both training and independent testing phases. It achieved a remarkably high mean F1-score of 0.996 during training and maintained strong generalization with a mean F1-score of 0.981 on the test set. This minimal performance drop indicates that FFBP2 successfully avoided overfitting, retaining its ability to generalize to new, unseen biomechanical data. Such stability is particularly valuable in real-world applications, where model performance must remain reliable across diverse populations and varying data conditions.

Notably, FFBP2 outperformed traditional machine learning classifiers such as Random Forest and Support Vector Machines not only in terms of raw performance metrics, precision, recall, accuracy, but also in terms of consistency and resilience to data variation. While RF and SVM showed competitive results, FFBP2 consistently maintained a better balance between sensitivity and specificity, which is essential in injury detection tasks where false negatives carry significant risk. Furthermore, FFBP2 exhibited lower standard deviation across repeated evaluations, underscoring its robustness and reliability as a decision-making tool.

Another key advantage of FFBP2 is its relatively simple architecture, which offers high performance without the computational complexity associated with deeper or more intricate models. This efficiency makes FFBP2 particularly suitable for implementation in resource-constrained environments, such as wearable sensor systems or real-time monitoring platforms, where computational overhead must be minimized without sacrificing accuracy.

Looking ahead, future work will focus on expanding the size and heterogeneity of the dataset to enhance the model's generalizability across a wider range of athletic profiles, injury types, and movement patterns. Incorporating more demographic and contextual variability will allow for a more comprehensive evaluation of the model's adaptability. Additionally, more sophisticated neural network architectures, such as convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs) or LSTMs for capturing temporal dependencies,will be explored to better model the dynamic nature of biomechanical data.

The integration of real-time sensor input streams and adaptive feedback mechanisms is also a priority, enabling the deployment of intelligent systems capable of continuous monitoring and timely intervention. Such systems could play a critical role in injury prevention and rehabilitation by providing actionable insights during athletic activity. Finally, future efforts will emphasize the standardization of preprocessing pipelines, careful documentation of hyperparameter settings, and improved interpretability through explainable AI methods, such as feature attribution and relevance mapping. These steps are essential for ensuring transparency, reproducibility, and trust in the deployment of machine learning solutions in health-related contexts.

ACKNOWLEDGMENT

The author expresses their deep gratitude to their thesis advisor, Noel Pérez, whose guidance and encouragement were instrumental to every phase of this work

BIBLIOGRAPHY

Rahlf, A., Hoenig, T., Stürznickel, J., Cremans, K., Fohrmann, D., Sanchez-Alvarado, A., Rolvien, T., & Hollander, K. (2022). A machine learning approach to identify risk factors for running-related injuries: Study protocol for a prospective longitudinal cohort trial. *BMC Sports Science, Medicine and Rehabilitation, 14*(1), 75.

Hancock, M. J. (2019). *Defining types and risk factors of overuse injuries of the foot, ankle, leg and knee in distance runners: A scoping review*, (Master's thesis).

Van Hooren, B., van Rengs, L., & Meijer, K. (2024). Predicting musculoskeletal loading at common running injury locations using machine learning and instrumented insoles. *Medicine & Science in Sports & Exercise, 56*(10), 2059–2075.

Taunton, J., Ryan, M., Clement, D., McKenzie, D., Lloyd-Smith, D., & Zumbo, B. (2003). A prospective study of running injuries: The Vancouver Sun Run "In Training" clinics. *British Journal of Sports Medicine, 37*(3), 239–244.

Seçkin, A., Ateş, B., & Seçkin, M. (2023). Review on wearable technology in sports: Concepts, challenges and opportunities. *Applied Sciences, 13*, 10399.

Bogaert, S., Davis, J., Van Rossom, S., & Vanwanseele, B. (2022). Impact of gender and feature set on machine-learning-based prediction of lower-limb overuse injuries using a single trunk-mounted accelerometer. *Sensors, 22*(8), 2860.

Wang, L. (2024). *Injury prediction in sports: A survey on machine learning methods*, .

Xiang, L., Gao, Z., Wang, A., Shim, V., Fekete, G., Gu, Y., & Fernandez, J. (2024). Rethinking running biomechanics: A critical review of ground reaction forces, tibial bone loading, and the role of wearable sensors. *Frontiers in Bioengineering and Biotechnology, 12*, 1377383.

Martínez-Gramage, J., Albiach, J. P., N. Moltó, I., Amer-Cuenca, J. J., Huesa Moreno, V., & Segura-Ortí, E. (2020). A random forest machine learning framework to reduce running injuries in young triathletes. *Sensors, 20*(21), 6388.

Alghamdi, W. Y. (2023). A novel deep learning method for predicting athletes' health using wearable sensors and recurrent neural networks. *Decision Analytics Journal, 7*, 100213. https://www.sciencedirect.com/science/article/pii/S277266222300053X

Leckey, C., Van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2025). Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. *British Journal of Sports Medicine, 59*(7), 491–500.

Li, W. (2024, September). A big data approach to forecast injuries in professional sports using support vector machine. *Mobile Networks and Applications* https://doi.org/10.1007/s11036-024-02377-x

Wu, X., Zhou, J., Zheng, M., Chen, S., Wang, D., Anajemba, J., Zhang, G., Abdelhaq, M., Alsaqour, R., & Uddin, M. (2022, November). Cloud-based deep learning-assisted system for diagnosis of sports injuries. *Journal of Cloud Computing, 11*(1), 82. https://doi.org/10.1186/s13677-022-00355-w

Murugan, S., Kumar, S. P., G, K., Saritha, B., & Rubini, B. (2025). Comparison of machine learning models for injury prediction in athletes. In *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, (pp. 107–112).

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics, 8*, 1–15.

Ferber, R., Brett, A., Fukuchi, R. K., Hettinga, B., & Osis, S. T. (2024). A biomechanical dataset of 1,798 healthy and injured subjects during treadmill walking and running. *Scientific Data, 11*(1), 1232.

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics, 81*, 1–11.

Pareek, A., Karlsson, J., Martin, R. K., et al. (2024). Machine learning/artificial intelligence in sports medicine: State of the art and future directions. *Journal of ISAKOS, 9*(4), 635–644.

Zadeh, A., Taylor, D., Bertsos, M., Tillman, T., Nosoudi, N., & Bruce, S. (2021). Predicting sports injuries with wearable technology and data analysis. *Information Systems Frontiers, 23*, 1023–1037.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology, 2*(4), .

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

John, G. H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964*, .

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(2), 111–133.

Tuychiev, B. (2022). F1 score vs accuracy vs ROC AUC vs PR AUC – Which metric is better? *Neptune.ai* https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Khera, P., & Kumar, N. (2020). Role of machine learning in gait analysis: A review. *Journal of Medical Engineering & Technology, 44*(8), 441–467.

Rezapour, M., Seymour, R. B., Sims, S. H., Karunakar, M. A., Habet, N., & Gurcan, M. N. (2023). Machine learning based analytics for the significance of gait analysis in monitoring and managing lower extremity injuries. *arXiv preprint*, .