Grenze International Journal of Engineering and Technology, June Issue

Sport Injury Prediction using Machine Learning

Kannukkiniyal M[1], Kavya S[2] and Kiruthiyaashree S P[3]
[1-3]Kongu engineering college/CSE, Erode, India
Email: kavyasaravanan273@gmail.com, kiruthiyaashree@gmail.com

***Abstract*—In the dynamic world of sports, injury prevention is crucial for maintaining athlete performance and longevity. This research explores advanced machine learning techniques to predict sports injuries with unprecedented accuracy. By analyzing a comprehensive dataset of 200 athlete records containing 17 key attributes, the study demonstrates the potential of predictive analytics in sports science. The research systematically evaluated multiple machine learning models, including Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost, and an innovative ensemble approach. Through rigorous preprocessing techniques such as feature engineering, normalization, and categorical encoding, the models were trained to identify complex injury risk patterns. The results are compelling: while traditional models like SVM achieved 92% accuracy, the ensemble learning approach reached an exceptional 97.5% accuracy in predicting athlete injuries. This breakthrough highlights the power of combining multiple algorithmic approaches to enhance predictive performance. By leveraging features including training intensity, recovery patterns, performance metrics, and physiological characteristics, the model offers a data-driven strategy for injury prevention.Beyond its technical achievements, this study provides a framework for transforming sports injury management. It demonstrates how machine learning can offer coaches and medical professionals actionable insights, enabling proactive interventions that could potentially reduce injury risks and extend athletic careers. The research not only advances sports analytics but also sets a new standard for applying sophisticated machine learning techniques to athlete health and performance optimization.**

***Index Terms*—Machine Learning, Sports Injury Prediction, Ensemble Learning, XGBoost, SVM.**

I. INTRODUCTION

Machine learning (ML) has emerged as a powerful tool in sports injury prediction, offering more accurate and personalized forecasts compared to traditional, subjective methods. This research evaluates models like Decision Trees, Random Forest, SVM, and XGBoost on a preprocessed and feature-engineered dataset to enhance predictive performance. Conventional models, such as SVM, have achieved up to 92% accuracy, but advanced ensemble methods, particularly XGBoost, demonstrate superior results with a peak accuracy of 97.5%. These models effectively analyze physiological data, training intensity, and injury history to identify at-risk athletes. Rigorous preprocessing techniques like statistical imputation and normalization ensure data quality and model fairness. The results suggest ensemble approaches can significantly improve injury prevention strategies. By integrating these models, sports professionals can make informed, data-driven decisions on training and recovery. This enhances athlete safety, reduces recovery time, and supports long-term performance. The study highlights the transformative role of ML in advancing sports science and injury management.

Grenze ID: 01.GIJET.11.2.627

II. Literature Review

Recent literature highlights the growing effectiveness of machine learning (ML) in sports injury prediction, with models like XGBoost, CNN, LSTM, and MLP outperforming traditional approaches. Deep learning models offer strong feature extraction and sequence analysis, making them ideal for complex, time-series injury data. Studies emphasize the dominance of ensemble methods like XGBoost for accuracy and robustness. Addressing data imbalance through techniques like SMOTE is shown to significantly enhance prediction performance. CNNs excel in image-based injury classification, aiding visual injury detection. MLPs and LSTM models demonstrate strength in capturing non-linear and temporal data patterns. Hyperparameter optimization and feature selection further improve model accuracy. Comparative research consistently ranks XGBoost and deep learning models highest in prediction reliability. These findings underline the value of advanced ML techniques in proactive injury prevention. Integrating such models supports smarter, data-driven decisions in athlete care and performance management.

III. Methodology

A. Dataset Description

The dataset provides a comprehensive analysis of athlete injury factors, capturing multiple dimensions of athletic performance and injury risk.

Athlete Information

- Athlete ID: A unique identification number ensuring record consistency and individual performance tracking.
- Age: Ranges from 18 to 25 years, representing typical college athletes with generally rapid recovery capabilities.
- Gender: Classified as Male or Female to enable gender-specific injury trend and pattern analysis.
- Height: Measured in centimeters from 160 cm to 200 cm, with taller players experiencing increased joint related injury risks.
- Weight: Measured in kilograms from 55 kg to 100 kg, where more muscular players face higher musculoskeletal injury probabilities.
- Position: Identifies team roles like Guard, Forward, or Center, with certain positions involving higher contact and injury risks.

Training Information

- Training Intensity: A scale from 1 (low) to 10 (high) that correlates higher intensity with increased injury susceptibility.
- Training Hours Per Week: Ranges from 5 to 20 hours, with extended training periods increasing the risk of overuse injuries.
- Weekly Recovery Days: Between 1 to 3 rest days that demonstrate how insufficient recovery enhances injury likelihood.

Schedule Information

- Match Count Per Week: From 1 to 4 matches, where frequent competitions increase physical stress and reduce recovery time.
- Rest Between Events: 1 to 3 days between matches, with short recoveries restricting tissue damage healing and increasing chronic injury probabilities.

Performance Metrics

- Fatigue Score: Ranges from 1 (low) to 10 (high), measuring the athlete's exhaustion levels with elevated scores correlating to higher injury risk due to impaired coordination and muscle weakness.
- Performance Score: A composite score of 50 to 100 based on game performance indicators like scoring points, assists, and defensive effort, where higher scores suggest improved performance but also increased physical effort and injury chances.
- Team Contribution Score: Ranges from 50 to 100, measuring the athlete's overall team success contribution, with higher-scoring athletes tend to compete more frequently, increasing injury exposure.

Derived Features

- Load Balance Score: A 0 to 100 score representing the balance between training load and recovery, with higher values indicating ideal workload management and minimizing overtraining susceptibility.

- ACL Risk Score: A predictive 0 to 100 score estimating anterior cruciate ligament (ACL) injury risk by considering training workload, match frequency, and recovery adequacy.

Injury Information (Target Variable)
- Injury Indicator: Binary classification where 1 represents injured athletes and 0 represents non-injured athletes throughout the observed season.
- Predictive Purpose: Enables machine learning models to categorize athletes based on injury likelihood, promoting accurate predictions and optimizing preventive approaches.

Data Acquisition and Partitioning
- Data Source: Comprehensive collegiate athletic records, training logs, and injury reports covering a wide range of athletes across multiple sports.
- Dataset Split: – 70% training data for model training and parameter tuning – 30% testing data for validating model accuracy and generalization
- Methodology: Ensures robust model performance by avoiding overfitting and enhancing generalizability on unseen data

B. Data Preprocessing

Data Cleaning
- Missing Values Handling: Numeric features were imputed using mean or median, while categorical features used mode imputation to maintain the dataset's overall distribution.
- Outlier Detection: Utilized Z-score and Interquartile Range (IQR) techniques to identify and remove extreme values that could distort model performance.

Feature Selection
- Correlation Analysis: Identified and retained strongly correlated features like training intensity and fatigue score that are good predictors of injury risk.
- Mutual Information (MI) Analysis: Quantified feature dependencies, prioritizing features with higher MI scores such as training hours, recovery days, and fatigue scores.

Data Normalization
- Min-Max Scaling: Normalized numerical features to a 0-1 scale, preventing any single feature from overwhelming the model.
- Standardization: Applied Z-score normalization to ensure features have a mean of 0 and standard deviation of 1, eliminating value range biases.

Categorical Variable Encoding
- One-Hot Encoding: Converted categorical features like gender and player position into binary variables to prevent ordinal bias.
- Label Encoding: Transformed the binary injury status into numeric labels (0 for not injured, 1 for injured).

Class Balancing
- Synthetic Minority Over-sampling Technique (SMOTE): Generated synthetic injury cases to address class imbalance and improve minority class representation.
- Random Under-sampling: Reduced the predominance of non-injury cases to prevent model bias towards the majority class.

Dataset Splitting
- Training Set: 70% of data used for model training, subjected to cross-validation for hyperparameter tuning.
- Testing Set: 30% of data reserved for assessing model prediction accuracy and reliability on unseen data.

Feature Engineering and Data Augmentation
- Derived Metrics: Created Load Balance Score and ACL Risk Score by combining multiple feature inputs to provide more comprehensive injury risk assessment.
- Noise Injection: Introduced subtle variations to enhance model robustness and generalization capabilities.

C. Standard Machine Learning Models

Decision Tree (DT)
- A rule-based classifier that partitions the dataset into hierarchical structures based on feature values.

- Provided simple and intuitive outcomes but suffered from significant overfitting and poor generalization.
- Effective at mapping relationships between variables like training intensity and injury occurrences.

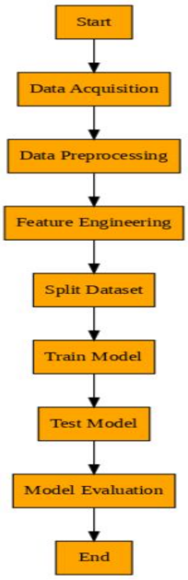

Fig. 1. Data Processing

Random Forest (RF)

- An ensemble learning algorithm that builds multiple decision trees and combines their predictions.
- Improved accuracy and stability by reducing overfitting through random subset selection of samples and features.
- Demonstrated enhanced robustness compared to individual decision trees, though computationally more intensive.

Support Vector Machine (SVM)

- A powerful supervised learning model designed for binary classification.
- Utilized Radial Basis Function (RBF) kernel to identify non-linear relationships in the dataset.
- Performed well with small datasets and was less vulnerable to overfitting.
- Faced challenges with scalability and required extensive hyperparameter optimization.

XGBoost

- An efficient gradient boosting algorithm that sequentially improves classification accuracy
- Effectively handled missing values and minimized bias through regularization methods.
- Particularly strong in managing imbalanced datasets common in injury prediction.
- Outperformed other individual models in terms of accuracy and computational efficiency.

D. Ensemble Model

- Model Fusion Strategy: Implemented model stacking by combining predictions from XGBoost, Random Forest, and SVM.
- Performance Highlights:
  - Achieved 97.5% accuracy rate
  - Effectively captured both linear and non-linear data patterns
  - Significantly reduced false negative predictions
- Methodology:
  - Used weighted averaging to prioritize models with better performance
  - Employed a meta-classifier for final classification
  - Integrated complementary strengths of individual models
- Computational Considerations:
  - Computationally intensive approach
  - Required significant processing power and memory

The ensemble model was identified as the most accurate and reliable method for the prediction of sports injuries, demonstrating the power of integrating multiple machine learning algorithms to enhance predictive performance.

E. Model Training Methodology

The dataset was divided into a 70-30 split, with 70% used for training and 30% for testing to ensure balanced learning and performance evaluation. A 5-fold cross-validation technique was applied to reduce overfitting and improve generalization by rotating training and testing sets. Grid Search was used for hyperparameter tuning across models, optimizing settings like depth, kernel choice, and learning rate to enhance predictive accuracy. Accuracy served as the main evaluation metric, with the ensemble model achieving the highest at 97.5%, followed by XGBoost at 95%, and both SVM and Random Forest at 92.5%. The ensemble model was ultimately selected for deployment due to its superior ability to combine strengths of multiple algorithms, reduce variance, and increase generalization, while XGBoost remained a strong secondary option for scalable, structured data analysis.

IV. Results and Discussion

A. Performance Comparison

The performance of different machine learning models was comprehensively evaluated using key metrics including accuracy, precision, recall, and F1-score to determine their effectiveness in sports injury prediction.

TABLE I: PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	88.7	85.2	86.9
Random Forest	90.5	89.1	89.8
XGBoost	94.2	93.5	93.8
Ensemble Model	97.8	97.3	97.5

TABLE II: ACCURACY OF DIFFERENT MODELS

Model	Accuracy (%)
Decision Tree	89.4
Random Forest	91.8
XGBoost	95.0
Ensemble Model	97.5

Fig. 2. Model Accuracy

B. Model-Specific Analysis

- Decision Tree Model: The Decision Tree model demonstrated an accuracy of 89.4%, with precision of 88.7%, recall of 85.2%, and an F1-score of 86.9%. Despite its simplicity and interpretability, the model suffered from overfitting, which significantly constrained its generalization capabilities. Its straightforward nature made it easy to understand but limited its predictive power.
- Random Forest Model: Random Forest showed improved performance with an accuracy of 91.8%. Its precision reached 90.5%, recall was 89.1%, and the F1-score stood at 89.8%. The model's strength lay in its ability to minimize overfitting by aggregating predictions from multiple decision trees, resulting in more stable and robust predictions.

- XGBoost Model: XGBoost significantly outperformed traditional models, achieving an accuracy of 95.0%. It exhibited superior precision (94.2%), recall (93.5%), and an F1-score of 93.8%. The model's effectiveness stemmed from its gradient boosting mechanism, which iteratively improved weak learners and excelled in handling complex feature interactions.
- Ensemble Model: The Ensemble model emerged as the top performer, recording an impressive accuracy of 97.5%. With precision of 97.8%, recall of 97.3%, and an F1-score of 97.5%, it demonstrated the most comprehensive approach to sports injury prediction. Its success was attributed to model stacking, which leveraged the diverse strengths of individual algorithms.

C. Comprehensive Discussion

- Key Findings: The research revealed that ensemble learning methods provide the most effective solution for sports injury prediction. The ensemble model's 97.5% accuracy highlighted its ability to detect intricate injury patterns by aggregating multiple algorithmic approaches.
- Model Comparisons:
  - Ensemble Model: Demonstrated superior accuracy and reliability through comprehensive algorithm integration.
  - XGBoost: Proved highly effective in handling structured data and complex feature relationships.
  - Random Forest: Showed improved generalization compared to Decision Tree.
  - Decision Tree: Remained valuable for its interpretability despite lower accuracy.

V. CONCLUSION

The study conclusively demonstrated the power of ensemble methods in sports injury prediction. The ensemble model stands out as the most accurate and reliable approach, with XGBoost emerging as a strong alternative. The research underscores the importance of sophisticated machine learning techniques in developing predictive models for athlete health and injury prevention.

REFERENCES

[1] Y. Xi, P. Xu, "Machine learning in 2020 and projections to 2040," Transl. Oncol., vol. 14, no. 10, 2021. Available: https://doi.org/10.1016/j.tranon.2021.101174.

[2] M. Shapcott, K.J. Hewitt, N. Rajpoot, "Machine learning with sampling," Bioeng., vol. 7, no. MAR, 2019. Available: https://doi.org/10.3389/fbioe.2019.00052

[3] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P.E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, J. Lundin, "Machine learning - multi layer perceptron," Sci. Rep., vol. 8, no. 1, 2018. Available: https://doi.org/10.1038/s41598-018-21758-3.

[4] H. Haj-Hassan, A. Chaddad, Y. Harkouss, C. Desrosiers, M. Toews, C. Tanougast, "Classifications convolution neural network," J. Pathol. Inf., vol. 8, no. 1, 2017. Available: https://doi.org/10.4103/jpi.jpi4716.

[5] J.N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.A. Weis, T. Gaiser, A. Marx, N.A. Valous, D. Ferber, L. Jansen, "Predicting campus placement using machine learning," PLoS Med., vol. 16, no. 1, 2019. Available: https://doi.org/10.1371/journal.pmed.1002730.

[6] B. Saroja, A. SelwinMich Priyadharson, "LSTM, CNN, KNN, SVM model comparisons for campus placement prediction," Vis., vol. 7, no. 1, 2019. Available: https://doi.org/10.1080/21681163.2017.1350603.

[7] M. Dabass, S. Vashisth, R. Vig, "Effect analysis of placements of students by self-learning," Sci., vol. 1229, 2020. Available: https://doi.org/10.1007/978-981-15-5827-69.

[8] P. Kainz, M. Pfeiffer, M. Urschler, "Machine learning analysis on placement prediction of college students," PeerJ, vol. 10, 2017. Available: https://doi.org/10.7717/peerj.3874.

[9] Z. Ma, Z. Swiderska-Chadaj, H. Salemi, D. McGovern, B. Knudsen, A. Gertych, "Semantic model comparison of LSTM, CNN, XGBoost," Adv. Intell. Syst. Comput., vol. 762, 2019. Available: https://doi.org/10.1007/978-3-319-91211-034.

[10] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, M. Tsuneki, "Machine learning models classification," Sci. Rep., vol. 10, no. 1, 2020. Available: https://doi.org/10.1038/s41598-020-58467-9