Comparative Study of Approaches for Injury Risk Prediction in Athletes

Team 27: Akanksh Rao, Anuj Joshi, Mansi Nandkar, Reshma Panibhate

Course: CSE 575 Statistical Machine Learning

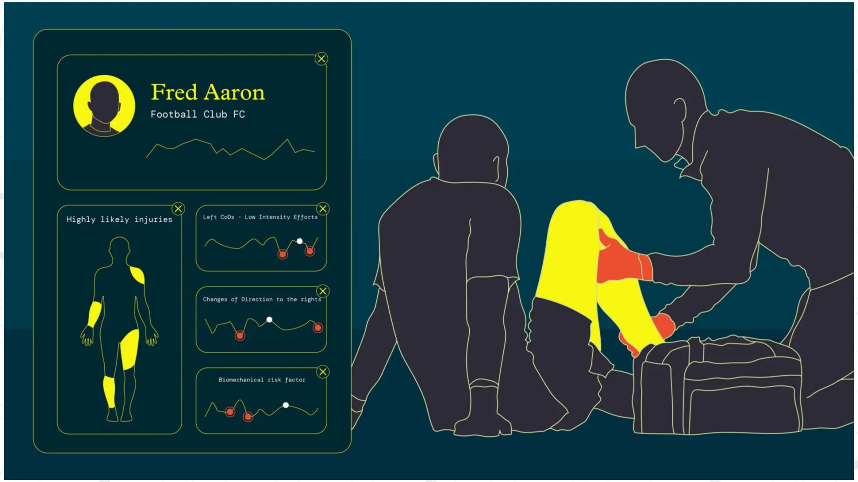
Date: 04/30/2025

Agenda

Introduction

- **Importance of Injury Prediction:**
  - Athlete injuries impact performance and careers.
  - Traditional methods lack real-time adaptability.

- **Role of ML/DL:**
  - Integration with wearable sensor data enables proactive injury risk assessment.



Literature Review

AI in Sports:

- Studies show **variable prediction performance**.
- Highlight need for **standardized evaluation metrics**.

Recent Advances:

- ML models using physiological and biomechanical variables show promise.
- RNN-based IoT systems enable **real-time injury prediction**.
- CNNs on wearable device data enhance **predictive performance and safety**.

Hybrid Approaches:

- Fusion of **data-driven methods with expert knowledge** improves robustness and interpretability.

Gaps Identified:

- Challenges remain in **model interpretability** and **scalability** across settings.

Problem Definition

- **Context**: Injury risk prediction in sports is a growing field but faces key challenges.
- **Challenges**:
  - Data quality issues
  - Poor model generalizability across different sports
  - Lack of interpretability in existing models
- **Gaps Identified**:
  - Inconsistent evaluation methods
  - Limited integration of domain knowledge into data-driven models
- **Our Objective**:
  - Conduct a **comparative analysis** of Machine Learning (ML) and Deep Learning (DL) approaches.
  - Improve prediction performance, model transparency, and practical applicability across varied settings.

Dataset Overview

MHEALTH (Mobile Health)

Dataset Source: UCI Machine Learning Repository
Data Collected From: 10 volunteers (8 male, 2 female)
Age Range: 20-35 years

Sensors Used:

- Accelerometer, Gyroscope, Magnetometer (Wrist + Ankle)
- Accelerometer, ECG (Chest)

Sampling Frequency: 50 Hz (50 measurements per second)

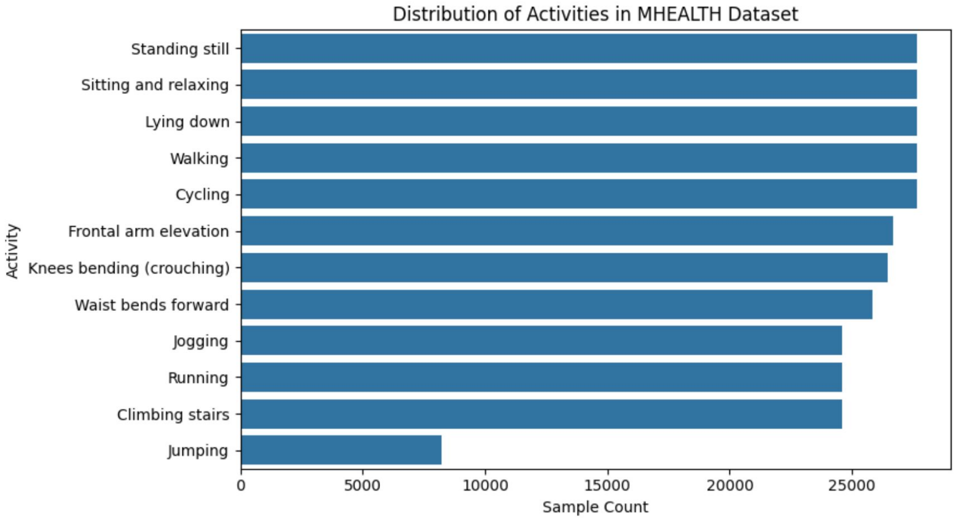
Features:

- 24 continuous sensor signals

Goal for our Project:

- Use this wearable sensor data to predict **injury risk**, not just activity type.

Activity Set



Risk Label Creation

- **Problem**: Dataset has no true injury labels.
- **Solution**: Create **Proxy Risk Labels** based on physical movement signals.
- **Indicators Used**:
    a. **High Impact Acceleration**:
        i. Chest acceleration > 3.5g
        ii. Indicates falls, unsafe landings
    b. **Fatigue Signals**:
        i. Elevated ECG heart rate during standing/sitting
        ii. Indicates cardiovascular strain
    c. **Repetitive Stress**:
        i. Long continuous dynamic activity (e.g., >150 steps without rest)
        ii. Simulates overuse injuries
    d. **Postural Instability**:
        i. High variance in gyroscope wrist signals
        ii. Indicates unstable body transitions
- **Final Risk Label**:
    a. Risk = 1 if any condition is triggered
    b. No Risk = 0 otherwise

Data Processing

Classical ML (Logistic Regression, Random Forest, Support Vector Machine)	LSTM	1D CNN
Windowing: 2s windows (100 samples) 50% overlap	**Windowing**: 2s windows (raw time-series) 50% overlap	**Windowing**: 2s windows (raw time-series) 50% overlap
Feature Extraction: Mean, Std, Max, Min, Energy, Peaks	**No Feature Extraction**: Use raw sequential data	**No Feature Extraction**: Use raw sequential data
Normalization: On extracted features	**Normalization**: StandardScaler on sensor features	**Normalization**: StandardScaler on sensor features
Balancing: **SMOTE** on feature vectors	**Balancing**: **Downsampling** No Risk samples	**Balancing**: **Downsampling** + **Weighted Loss**
Split: Subject-wise Split	**Split**: Subject-wise Split	**Split**: Subject-wise Split

Model Selection

- **Machine Learning Models:**
    - Logistic Regression
    - Random Forest
    - Support Vector Machine

    Reasons:

    i. Fast, interpretable baselines
    ii. Handles small feature sets well
    iii. Helps evaluate statistical vs deep approaches

Model Selection

- **Deep Learning Models:**
  - **Long Short-Term Memory Network (LSTM) (sequential)**

    Reasons:

    i. Captures long-term dependencies

    ii. Learns from full time-series

    iii. Suited for sequential patterns

  - **1D Convolutional Neural Network (1D CNN) (local temporal)**

    Reasons:

    i. Learns local signal variations

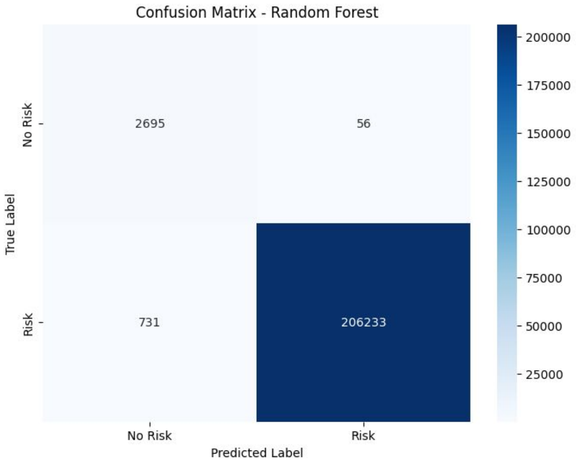
    ii. Lightweight & efficient

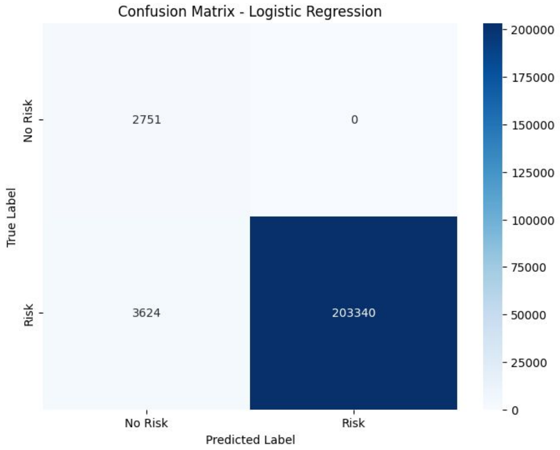
    iii. Strong performance with time-series data

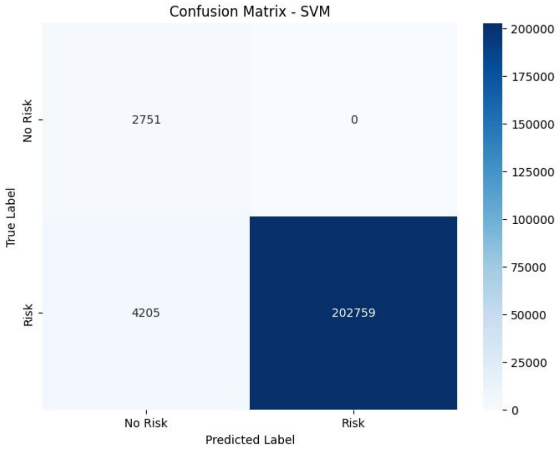
Evaluation Matrices

MODELS	ACCURACY	PRECISION	F1 SCORE	RECALL
Random Forest	100.00	100.00	100.00	100.00
Logistic Regression	98.00	100.00	99.00	98.00
Support Vector Machine	98.00	100.00	99.00	98.00
1D CNN	99.77	99.72	99.79	99.86
LSTM	90.82	88.69	92.04	95.65

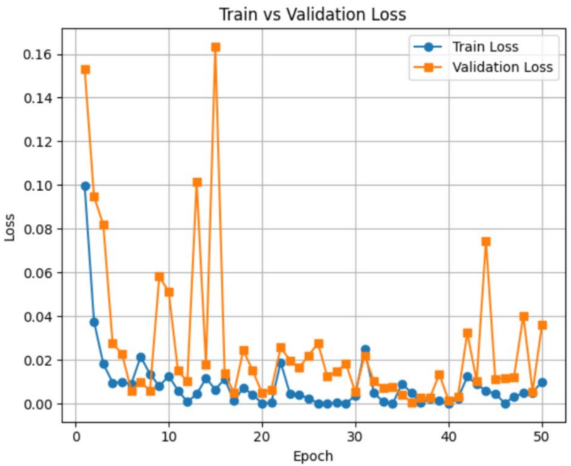
Visualisation of Results - ML Models

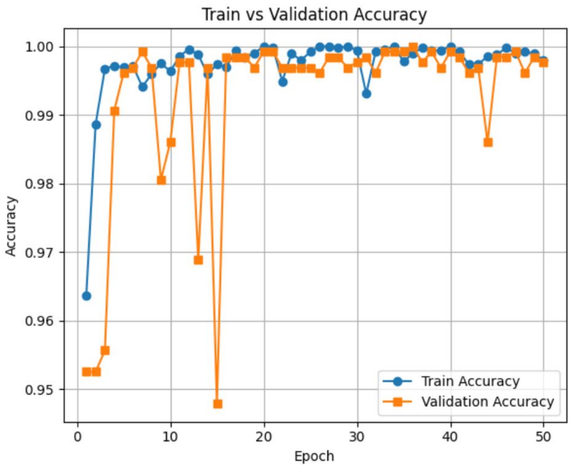


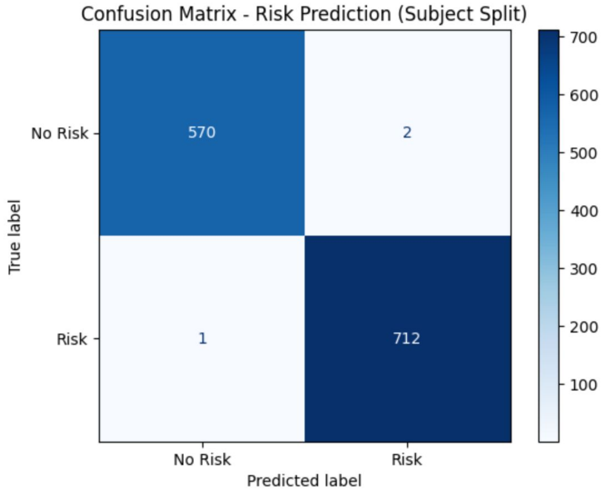


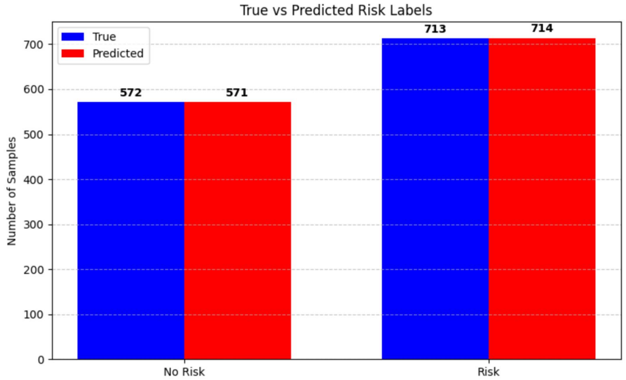


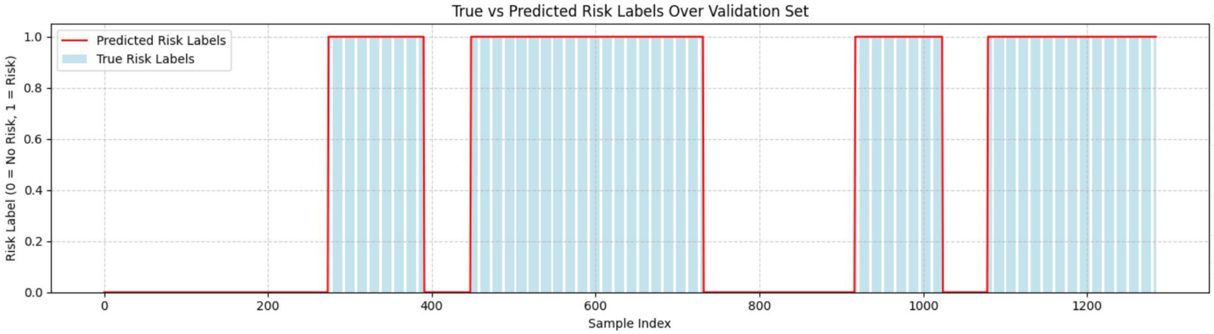
Visualisation of Results - 1D CNN



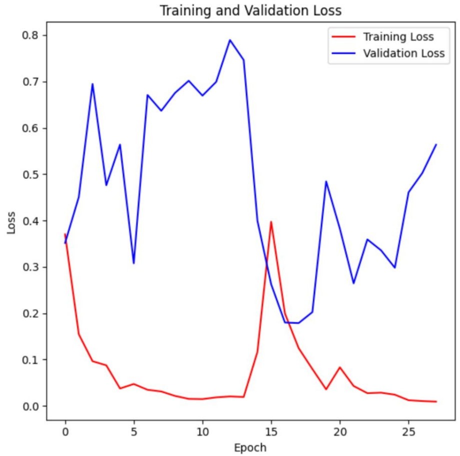


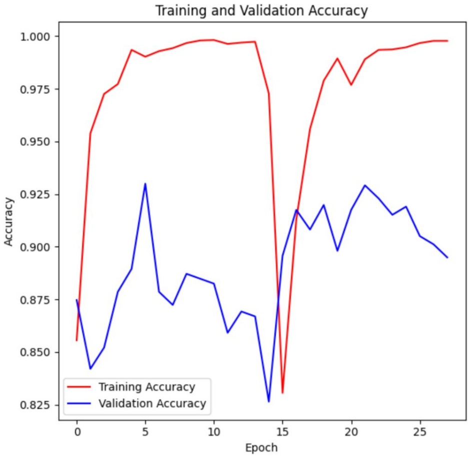


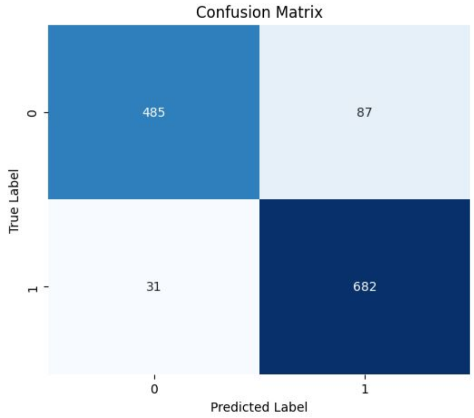


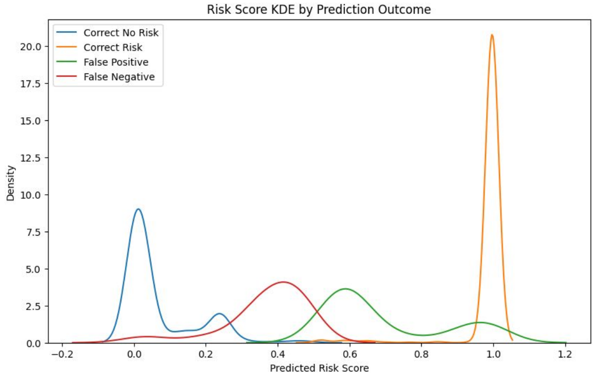


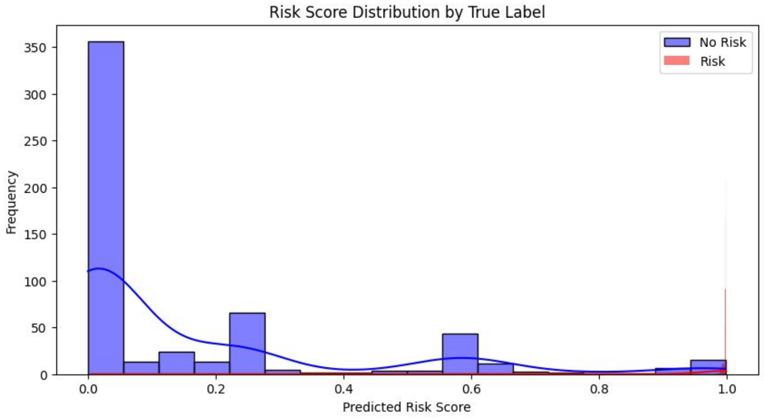
Visualisation of Results- LSTM











Insights and Takeaways

Classical ML Models (Random Forest, SVM, Logistic Regression)

- Achieved **near-perfect performance** (Accuracy ~98–100%, F1 Score ~99–100%).
- Likely benefited from well-separated, feature-engineered inputs and balanced datasets via SMOTE.
- Random Forest achieved 100% across all metrics may suggest **overfitting** or exceptionally clean decision boundaries.

1D CNN

- Scored 99.77% accuracy, 99.72% precision, and 99.86% recall.
- Very high F1 score (99.79) confirms excellent balance of precision and recall.
- Proves that 1D CNN can **extract and learn meaningful patterns** directly from raw sensor data.
- **Strong candidate for real-time** or embedded applications due to its efficiency

LSTM

- Lower precision (88.69%) compared to other models.
- Best recall (95.65%) **great at detecting risky windows**, but more false positives than CNN.
- Shows strength in capturing temporal dependencies, but may be affected by training duration or data variance.

Conclusion & Future Work

Summary:

The comparative analysis underscores the effectiveness of models like Random Forest and SVM in injury risk prediction, with strengths in accuracy, recall, and precision. The integration of wearable sensor data with ML/DL models offers significant potential for proactive risk prediction, enabling real-time insights and better training optimization.

Future Work:

Continued Research:

Further research is necessary to refine these models and improve their generalization, especially when dealing with real-world data or explore advanced models like Transformers for sequential data.

Collaboration with Sports Professionals:

Collaborating with sports scientists and health professionals will help tailor these models for practical, real-world use and ensure they are addressing the key injury predictors effectively.

THANK YOU