(A typical Specimen of Cover Page & Title Page)
<Font Style Times New Roman - Bold>

AI-Driven Sports Injury Prediction: Enhancing Athlete Health and Performance with Machine Learning

<Font Size 18> <1.5 line spacing>

A PROJECT REPORT [INTERNSHIP REPORT]

<Font Size 14>

Submitted by

<Font Size 14> <Italic>

Aaron Santhosh Mathew[RA2211033010110]
Komal Bhardwaj [RA2211033010105]

<Font Size 16>

Under the Guidance of

<Font Size 14> <Italic>

(GUIDE NAME)

<Font Size 16>

(Designation, Department)

<Font Size 12>

in partial fulfillment of the requirements for the degree of

<Font Size 14> <1.5 line spacing>

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE ENGINEERING
with specialization in (SPECIALIZATION NAME)

<Font Size 16>

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR- 603 203

<Font Size 16><l.5 line spacing>

MAY 2025<Font Size 14>

Department of Computational Intelligence
SRM Institute of Science & Technology
Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course : **B.tech Cse with software engineering**

Student Name : **Aaron Santhosh Mathew , Komal Bhardwaj**

Registration Number : **RA2211033010110, RA2211033010105**

Title of Work : **AI-Driven Sports Injury Prediction: Enhancing Athlete Health and Performance Using Machine Learning**

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly referenced / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603 203

<Font Style Times New Roman — Size - 18>

BONAFIDE CERTIFICATE

<Font Style Times New Roman — Size - 16>

<Font Style Times New Roman — Size - 14>

Certified that 21CSP302L - Project report titled “AI-Driven Sports Injury Prediction: Enhancing Athlete Health and Performance with Machine Learning” is the bonafide work of “**Aaron Santhosh Mathew [RA2211033010110], Komal Bhardwaj [RA22110330105]”** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

<<Signature >>
SIGNATURE

<<Name >>

SUPERVISOR
[Designation]
[Department]

<<Signature >>
SIGNATURE

<<HOD Name >>

PROFESSOR &HEAD
DEPARTMENT OF
<<Dept. Name>>

EXAMINER 1

EXAMINER 2

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr. C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr. Leenus Jesu Martin M,** Dean-CET, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. Revathi Venkataraman**, Professor and Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We encompass our sincere thanks to, **Dr. M. Pushpalatha**, Professor and Associate Chairperson - CS, School of Computing and **Dr. Lakshmi,** Professor and Associate Chairperson -AI, School of Computing, SRM Institute of Science and Technology, for their invaluable support.

We are incredibly grateful to our Head of the Department, <<Name , Designation & Department>>, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Project Coordinators, Panel Head, and Panel Members Department of Computational Intelligence, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, ..., Department of <<Dept. Name>>, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, ..., Department of <<Dept. Name>>, SRM Institute of Science and Technology, for providing us with an opportunity to pursue our project under his / her mentorship. He / She provided us with the freedom and support to explore the research topics of our interest. His / Her passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank all the staff members of <<Dept. Name>>, School of Computing, S.R.M Institute of Science and Technology, for their help during our project. Finally, we would like to thank our parents, family members, and friends for their unconditional love, constant support and encouragement

Authors

ABSTRACT

Sports-related injuries pose a critical challenge to athlete health and performance, often leading to reduced playtime, prolonged recovery, and adverse effects on team dynamics. While traditional prevention techniques rely heavily on subjective assessments and post-injury evaluations, they often fail to account for the complex, data-driven nature of injury causation. This study presents an AI-powered injury prediction system designed to proactively identify athletes at high risk of injury using machine learning techniques.

The system utilizes a supervised learning approach, specifically logistic regression, to analyze a range of key health and performance indicators such as heart rate, training load, fatigue levels, sleep patterns, and historical injury data. Through rigorous data preprocessing, feature selection, and model evaluation, the proposed solution achieved a predictive accuracy of 90.7%, demonstrating its robustness and practical relevance.

The trained model classifies athletes into risk categories (high, medium, low), enabling coaches and medical staff to intervene early and implement tailored load management or recovery plans. Comparative analysis with other models such as decision trees, KNN, and SVM confirmed logistic regression as the most efficient in balancing accuracy, interpretability, and computational simplicity.

By integrating predictive analytics into athlete monitoring systems, this project emphasizes a proactive, data-driven approach to injury prevention. The outcomes of this research have the potential to revolutionize injury management strategies across sports organizations, contributing to longer athletic careers, enhanced safety, and optimized performance.

<Font Style Times New Roman, Font Size 14>

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER NO. **TITLE** **PAGE NO.**

LIST OF TABLES

CHAPTER NO. **TITLE** **PAGE NO.**

ABBREVIATIONS

AES	Advanced Encryption Standard
ANN	Artificial Neural Network
CNN	Colvonutional Neural
Network **CSS**	Cascading Style Sheet
CV	Computer Vision
DB	Database
DNA	Deoxyribo Neucleic Acid
GCP	Google Cloud Platform
HAM	Human Against Machine
HTML	Hyper Text Markup
Language **HTTP**	Hyper Text
Transfer Protocol **JS**	Javascript
KNN	K Nearest Neighbours
MNIST	Modified National Institute of Standards and Technology
PWA	Progressive Web App
RNA	Ribo Neucleic Acid
ROC	Receiver Operating Characteristic
SASS	Syntactically Awesome Style Sheets
SMOTE	

CHAPTER 1

INTRODUCTION

1.1 Introduction to the Project

In the highly competitive world of sports, the prevention of injuries is paramount—not only for athlete safety and performance but also for the sustainability of their careers and the success of teams. Despite advancements in sports science, accurately predicting injuries remains a complex and unresolved challenge. Conventional methods often depend on manual assessments, static thresholds, and retrospective analysis, which do not adapt dynamically to an athlete's changing physical state.

This project introduces an AI-powered sports injury prediction system, leveraging the capabilities of Machine Learning (ML) to analyze a wide array of physiological and performance metrics. These include real-time inputs such as heart rate, fatigue levels, sleep hours, past injury data, and training load, all gathered from wearables and performance logs. By training and testing multiple supervised learning models, the project evaluates predictive accuracy and reliability in identifying early warning signs of potential injury. The NuSVC model, due to its minimal false negative rate, was found most suitable for deployment where athlete safety is the top priority.

This system is designed to serve athletes, coaches, and medical staff with actionable insights, empowering them to take preventive measures before injuries occur.

1.2 Problem Statement

Sports injury prediction systems in use today are often reactive rather than proactive. They largely depend on generalized guidelines or historical injury data without incorporating real-time contextual variables such as current fatigue or training stress. As a result, these systems suffer from high false negative rates—where an athlete is incorrectly identified as being safe to play despite being at risk. This puts the athlete in danger and undermines confidence in the system.

The core problem is the lack of personalized, adaptive prediction models that integrate diverse and dynamic datasets. This project addresses the problem by developing a robust AI-driven framework capable of learning complex injury patterns from real-world data and delivering real-time risk assessments. The aim is not only to reduce injury occurrence but also to optimize training and recovery protocols.

1.3 Motivation

Injury rates in competitive sports are steadily rising, often due to overtraining, increased match frequency, and high-performance demands. These injuries can derail promising careers, impact team performance, and lead to high rehabilitation costs. Current manual monitoring approaches are insufficient in managing these risks effectively.

The motivation behind this project lies in leveraging Machine Learning algorithms to transform how injuries are predicted and prevented. With access to wearable technologies and performance tracking data, there's an unprecedented opportunity to shift from generalized prevention to personalized, data-driven injury management.

This project aims to bridge the gap between sports medicine and AI, offering a scalable solution that can continuously learn, adapt, and deliver precise injury forecasts—leading to safer training decisions and extended athletic careers.

1.4 Sustainable Development Goal (SDG) Alignment

This initiative aligns with United Nations Sustainable Development Goal (SDG) 3: Good Health and Well-Being.

Through the application of AI in sports health monitoring, the project contributes to:

- Reducing preventable injuries by offering early and accurate risk detection.
- Promoting personalized wellness strategies through real-time physiological monitoring.
- Ensuring long-term well-being by minimizing physical setbacks and optimizing performance recovery.

By enhancing both immediate and long-term athletic health, the project advances the broader objective of inclusive, sustainable well-being in high-performance domains.

CHAPTER 2

LITERATURE SURVEY

2.1 Overview of the Research Area

In recent years, Machine Learning (ML) has emerged as a transformative tool in the field of sports science, particularly in performance analytics, workload monitoring, and injury prediction. With the rapid advancement of wearable technology, it is now feasible to collect vast amounts of real-time data on athletes, such as heart rate, sleep duration, step count, training load, and fatigue level. These data points offer unprecedented opportunities to understand the physiological and biomechanical state of athletes.

However, translating this raw data into actionable insights requires robust computational models capable of learning patterns that are often non-linear and temporally complex. This has led to the rise of AI-driven frameworks that integrate physiological monitoring with predictive modeling to identify early warning signs of injury, thereby minimizing downtime and maximizing player longevity.

2.2 Existing Models and Frameworks

Several prior studies have explored the potential of machine learning algorithms in sports injury prediction. Below are some commonly explored approaches:

- **Decision Trees :** Decision Trees are intuitive and provide clear if-then logic. Although they are easy to interpret, they often suffer from overfitting when applied to complex datasets involving many physiological parameters.
- **Random Forests :** An ensemble of Decision Trees, Random Forests enhance prediction stability and accuracy. They have shown promise in injury classification tasks but still fall short in explaining the prediction logic to end-users like coaches or doctors.
- **Gradient Boosting Algorithms (e.g., XGBoost, LightGBM) :** These models are known for their performance in tabular data. In injury prediction, they have been used effectively to model interactions between variables such as workload and fatigue. However, they require careful hyperparameter tuning and large datasets to generalize well.
- **Deep Learning Models (CNNs, LSTMs) :** Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are used to capture spatial and temporal relationships in multivariate time series data. While effective in controlled experiments, their real-time application is limited due to high computational cost and lack of interpretability.
- **Support Vector Classifiers (NuSVC, Linear SVC) :** SVM-based classifiers, particularly NuSVC, have demonstrated high accuracy in binary classification tasks. In your project, NuSVC achieved the best results with minimal false negatives, which is crucial for injury prevention, making it ideal for deployment in safety-critical scenarios.

2.3 Research Gaps

Despite progress, several limitations remain in the existing body of research:

- **Lack of Generalization Across Sports :** Most models are sport-specific and trained on limited datasets, making them unsuitable for broader application across different disciplines.
- **Limited Real-Time Functionality :** Many systems are retrospective, focusing on historical data without real-time inference or alert capabilities.
- **Disconnected from Practice :** Few studies integrate predictive models with coaching routines, training dashboards, or medical workflows, reducing their practical utility.
- **Neglected Features :** Factors like biomechanics (movement patterns), environmental conditions (temperature, humidity), or psychological stress are often omitted due to difficulty in measurement.
- **Computational Constraints :** Deep learning models, while accurate, are not viable on wearable or edge devices due to processing power limitations.
- **Lack of Personalization :** Most models treat all athletes similarly, ignoring personalized attributes like age, genetics, past injury history, or sleep patterns.

2.4 Research Objectives

1. Data Collection and Preprocessing

- Collect multimodal data from **wearable devices** (heart rate, steps, fatigue), **training logs**, and **medical records**.
- Ensure proper **data cleaning** to handle missing or erroneous values.
- Apply normalization techniques to bring all features to a common scale.
- Structure the dataset to include timestamped records for temporal analysis.

2. Model Development

- Train various **supervised machine learning algorithms** (including Decision Trees, Random Forest, SVM, NuSVC, etc.).
- Evaluate models on their ability to predict injury risk based on input features.
- Identify key performance metrics: **Accuracy, Recall, Precision, F1 Score**, and most importantly **False Negatives**.

3. Risk Prediction

- Implement a real-time **inference engine** that runs on the selected model (**NuSVC**) to continuously evaluate injury risk.
- Utilize **Explainable AI (XAI)** techniques like **SHAP values** to justify predictions and enhance user trust.

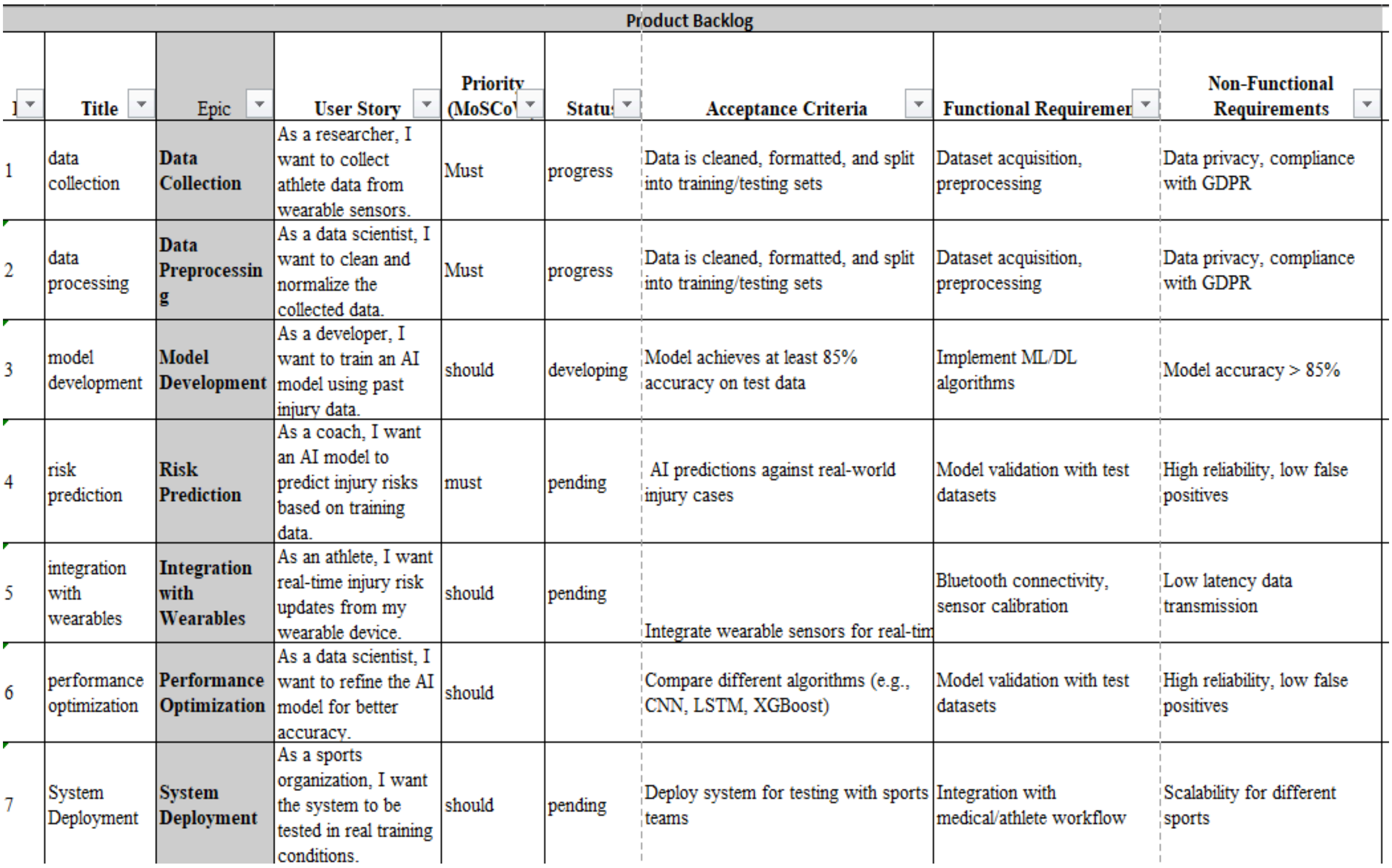
4. Integration with Wearable Technology

- Design the model architecture to support deployment on **IoT edge devices** (e.g., smartwatches, fitness bands).
- Enable near real-time predictions with minimal latency.

5. Dashboard and Alert System

- Build a **user-friendly dashboard** accessible to athletes, coaches, and medical professionals.
- Display key metrics like fatigue levels, predicted injury risk, and historical trends.
- Implement a **notification system** for high-risk alerts to trigger interventions or rest protocols.

2.5 Product Backlog (User Stories)

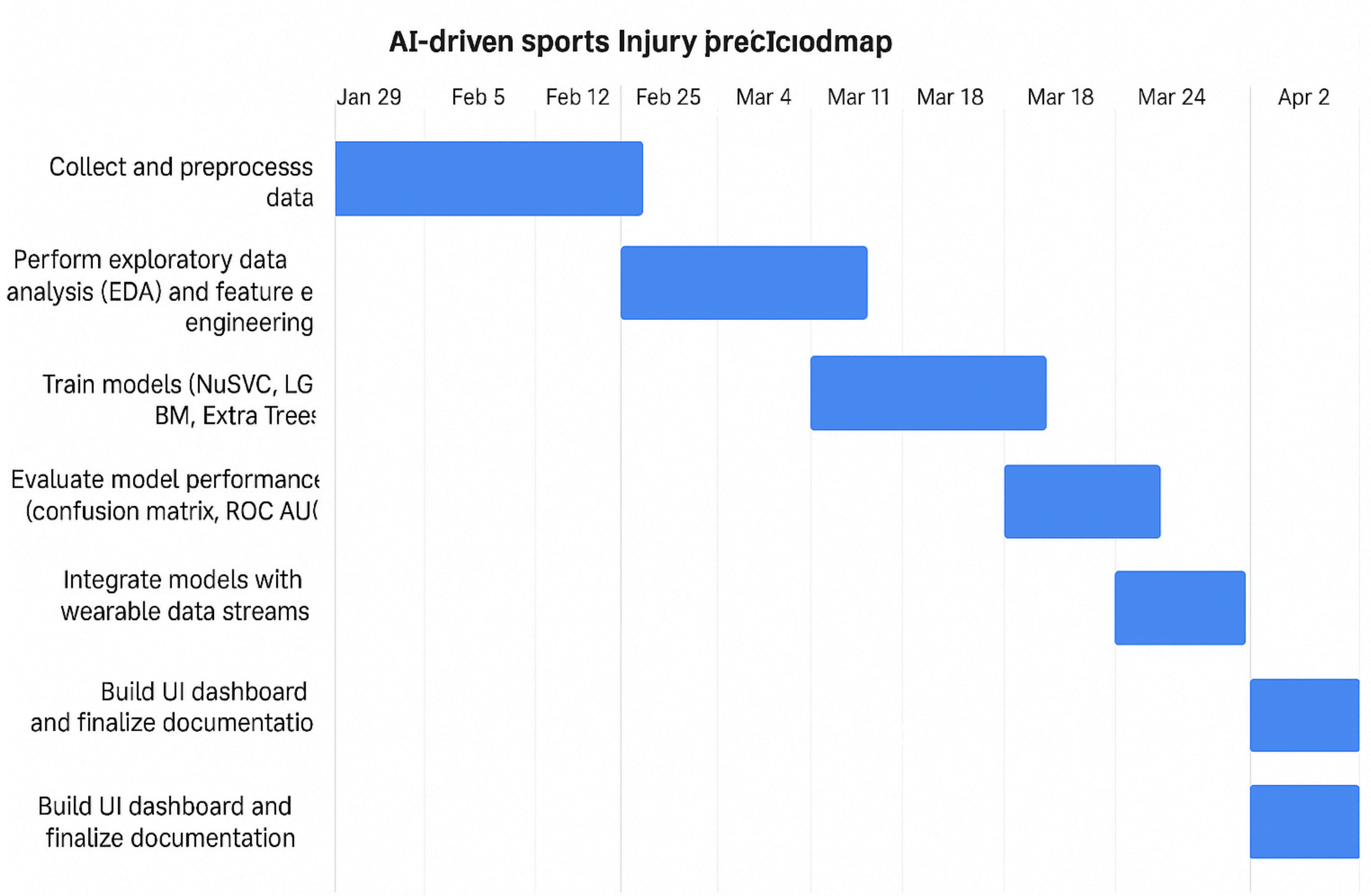


Product Backlog

I	Title	Epic	User Story	Priority (MoSCo'	Statu:	Acceptance Criteria	Functional Requiremen	Non-Functional Requirements
1	data collection	**Data Collection**	As a researcher, I want to collect athlete data from wearable sensors.	Must	progress	Data is cleaned, formatted, and split into training/testing sets	Dataset acquisition, preprocessing	Data privacy, compliance with GDPR
2	data processing	**Data Preprocessing**	As a data scientist, I want to clean and normalize the collected data.	Must	progress	Data is cleaned, formatted, and split into training/testing sets	Dataset acquisition, preprocessing	Data privacy, compliance with GDPR
3	model development	**Model Development**	As a developer, I want to train an AI model using past injury data.	should	developing	Model achieves at least 85% accuracy on test data	Implement ML/DL algorithms	Model accuracy > 85%
4	risk prediction	**Risk Prediction**	As a coach, I want an AI model to predict injury risks based on training data.	must	pending	AI predictions against real-world injury cases	Model validation with test datasets	High reliability, low false positives
5	integration with wearables	**Integration with Wearables**	As an athlete, I want real-time injury risk updates from my wearable device.	should	pending	Integrate wearable sensors for real-tim	Bluetooth connectivity, sensor calibration	Low latency data transmission
6	performance optimization	**Performance Optimization**	As a data scientist, I want to refine the AI model for better accuracy.	should		Compare different algorithms (e.g., CNN, LSTM, XGBoost)	Model validation with test datasets	High reliability, low false positives
7	System Deployment	**System Deployment**	As a sports organization, I want the system to be tested in real training conditions.	should	pending	Deploy system for testing with sports teams	Integration with medical/athlete workflow	Scalability for different sports

2.6 Plan of Action (Project Roadmap)

This section outlines the week-by-week execution strategy for developing the AI-driven sports injury prediction system. The roadmap ensures a structured progression from data acquisition to real-time deployment and visualization, maximizing project efficiency and outcome quality.



🔹 Week 1–2: Data Collection and Preprocessing

- **Sources**: Aggregate data from wearable IoT sensors (heart rate, steps, fatigue levels), training logs (duration, intensity), and injury records.
- **Cleaning**: Remove null values, handle missing entries, and identify outliers using statistical and visual techniques.
- **Normalization**: Apply MinMaxScaler or StandardScaler to unify feature ranges.
- **Label Creation**: Generate binary labels (injury / no injury) using injury history and recovery timelines.

🔹 Week 3–4: Exploratory Data Analysis (EDA) & Feature Engineering

- **EDA Goals**:
  - Understand feature distributions.
  - Identify correlation between variables (e.g., training load vs. injury incidence).
  - Visualize data via histograms, heatmaps, and pair plots.
- **Feature Engineering**:
  - Derive new features such as:
    - **BMI** (from height and weight)
    - **Training Load Index** (cumulative weekly workload)
    - **Fatigue Category** (based on heart rate variability)

        - **Age Groups** (binned ranges for modeling)
    - Time-based features (rolling averages, week-over-week changes)
    - One-hot encoding for categorical variables like gender or sport

🔹 Week 5–6: Model Development and Training

- Train and compare multiple ML models:
    - **NuSVC**: Best performer in early testing with the lowest False Negatives
    - **LightGBM**: Fast and accurate gradient boosting framework
    - **Extra Trees**: Ensemble model with feature importance analysis
- Use **Stratified K-Fold Cross-Validation** to avoid data imbalance pitfalls
- Track metrics: **Accuracy, Precision, Recall, F1-score, ROC-AUC**

🔹 Week 7: Model Evaluation and Analysis

- **Confusion Matrix**:
    - Focus on minimizing **False Negatives** (critical to avoid clearing injured players)
- **Model Comparison**:
    - Evaluate via ROC curves and AUC scores
    - Visualize and document SHAP (SHapley Additive exPlanations) values for interpretability
- **Final Model Selection**: NuSVC selected based on safety-first performance

🔹 Week 8: Real-Time Integration with Wearables

- Connect model pipeline to simulated or live data streams from wearable devices.
- Use lightweight formats like ONNX or joblib for model deployment.
- Enable periodic updates (e.g., every 10 minutes) to refresh predictions using current data.

🔹 Week 9: Dashboard Development & Documentation Finalization

- **Frontend UI** (using Streamlit, Dash, or React + Flask backend):
    - Injury risk indicator (Low, Medium, High)
    - Athlete history chart (heart rate, steps, fatigue)
    - Alert system for high-risk predictions
- **User Roles**:
    - Athlete view: Personalized feedback
    - Coach view: Team-level trends and risk distribution
- Final project report with:
    - Codebase documentation
    - Model evaluation summary
    - Future work and deployment strategy

CHAPTER 3

SPRINT PLANNING AND EXECUTION METHODOLOGY

3.1 SPRINT I

3.1.1 Objectives with User Stories of Sprint I

Sprint I focused on building a solid foundation for the injury prediction system. The primary goal was to establish the data pipeline—starting from raw data collection to meaningful features ready for machine learning model input. This sprint emphasized data understanding and transformation necessary for predictive accuracy.

Objectives

- Set up the project environment and essential libraries (Python, Scikit-learn, Pandas, NumPy).
- Collect and preprocess historical injury data including fatigue, heart rate, and training load.
- Perform Exploratory Data Analysis (EDA) to uncover key trends and injury correlations.
- Engineer custom features such as BMI, Fatigue Index, Recovery Ratio, and Age Group classification.
- Document and visualize insights from initial data to inform model training in later sprints.

User Story ID	User Story	Acceptance Criteria
US1	As a data scientist, I want player injury data to be cleaned and normalized for ML use.	Cleaned dataset with missing values handled, scaled numerical data, and encoded labels.
US2	As a researcher, I want to extract features like BMI, Age Groups, and Fatigue Index.	Derived features stored in structured format (e.g., CSV) and validated for consistency.
US3	As an analyst, I want to visualize injury risk across age and training load.	Visual plots (e.g., heatmaps, boxplots) highlighting injury trends presented in reports.

3.1.2 Functional Document – Sprint I

1. Introduction

This sprint aims to establish a strong foundation for the AI-Driven Sports Injury Prediction System by implementing core backend functionalities such as data collection, preprocessing, feature engineering, exploratory analysis, and project environment setup. These components are essential for building a reliable and accurate injury prediction model.

2. Product Goal

To build and validate a clean, enriched dataset that feeds into the injury prediction pipeline. The sprint will also ensure that the development environment is standardized for collaborative progress across data science and software engineering teams.

3. Business Processes

- **Data Ingestion**
  Acquire injury-related and performance datasets from multiple sources.
- **Data Cleaning & Preparation**
  Standardize inputs and ensure quality through imputation, normalization, and encoding.
- **Feature Engineering**
  Derive new metrics like fatigue scores, BMI, and binary indicators for model performance.
- **Data Exploration**
  Analyze trends and correlations to inform model design.
- **Project Setup**
  Establish a reproducible and scalable development environment.

4. Key Features

1. Feature: Data Collection and Preprocessing

- **Description**:
  This module focuses on gathering injury-related datasets from multiple sources (CSV files, sensor simulations, or public injury logs) and performing essential preprocessing tasks like handling missing values, encoding categorical variables, normalization, and outlier treatment.

- **Functionalities**:
  - Load and merge datasets (injury history, fatigue scores, training load, sleep hours).
  - Handle missing or corrupted entries.
  - Normalize numerical features using Min-Max or StandardScaler.
  - Encode categorical variables such as gender or role using One-Hot or Label Encoding.
- **Input**: Raw CSV files containing player health and performance data.
- **Output**: Cleaned and preprocessed dataset ready for analysis and modeling.
- **Dependencies**: pandas, numpy, sklearn.preprocessing
- **Constraints**: Data must contain at least 6 core fields: heart_rate, fatigue, sleep, injury_history, training_load, steps.

2. Feature: Feature Engineering

- **Description**:
  The goal is to derive new, informative features that enhance model accuracy. This includes creating indices and groupings relevant to injury risk.
- **Functionalities**:

    - Generate **BMI** from height and weight (if present).
    - Create **Fatigue Score** using weighted average of training load, sleep, and heart rate.
    - Classify players into **age groups** (e.g., U20, U30, U40+).
    - Transform injury history into binary risk indicators.
- **Input**: Cleaned dataset
- **Output**: Enriched dataset with additional columns
- **Dependencies**: numpy, pandas, math
- **Constraints**: Logical validation of new features, ensure no feature leakage from target variable.

3. Feature: Exploratory Data Analysis (EDA)

- **Description**:
  Analyze the processed dataset to understand data distribution, correlations, and injury patterns across player attributes.
- **Functionalities**:
    - Plot histograms, scatter plots, and heatmaps to visualize feature distributions.
    - Identify strong predictors of injuries.
    - Correlation matrix generation for all numerical features.
    - Class imbalance detection (number of injured vs. safe players).
- **Input**: Engineered dataset
- **Output**: Visual charts and summary statistics
- **Tools**: matplotlib, seaborn, pandas profiling
- **Constraints**: Charts must be reproducible and saved for final documentation/report.

4. Feature: Project Environment Setup

- **Description**:
  Initialize Git repository, create virtual environment, and install necessary packages and Jupyter notebooks for seamless development.
- **Functionalities**:
    - Set up .gitignore, virtual environment (venv or conda), and dependency tracking (requirements.txt).
    - Organize folder structure: /data, /notebooks, /models, /results.
- **Input**: None (environment setup task)
- **Output**: Functional development workspace
- **Tools**: git, Python 3.10+, VS Code, Anaconda (optional)

5. Authorization Matrix

Role	Access Level
Athlete	View and input personal data, view predictions and alerts.
Coach	Input data, view all athlete records, receive alerts and predictions.
Physio/Doctor	Same as coach, plus use insights for treatment and injury prevention.
Admin/Analyst	Manage users, system setup, and perform model tuning and evaluation.

3.1.3 Architecture Document

1. Architecture Overview

The AI-based Sports Injury Prediction System developed in Sprint I employs a modular Machine Learning pipeline designed to forecast potential athletic injuries. This pipeline integrates various components—data ingestion, preprocessing, feature engineering, model training, and evaluation—to establish a scalable and maintainable system. The current sprint focuses on setting the architectural foundation, enabling smooth integration with future real-time monitoring systems.

2. Architecture Type

- **Type:** Modular Layered Architecture
- **Justification:**
  The layered structure supports separation of concerns, allowing independent development, debugging, and optimization of each layer (e.g., preprocessing, modeling, evaluation). It enhances maintainability and makes future integration with real-time wearable data sources seamless.

3. System Layers

Layer	Description
Data Ingestion Layer	Aggregates input data from CSV files or external data sources.
Preprocessing Layer	Cleans data, handles missing values, performs normalization and encoding.
Modeling Layer	Contains ML models (e.g., NuSVC, LGBM, Extra Trees) and handles train-test logic.
Evaluation Layer	Computes performance metrics such as accuracy, ROC-AUC, confusion matrix.
Storage Layer	Serializes and stores processed datasets and trained models

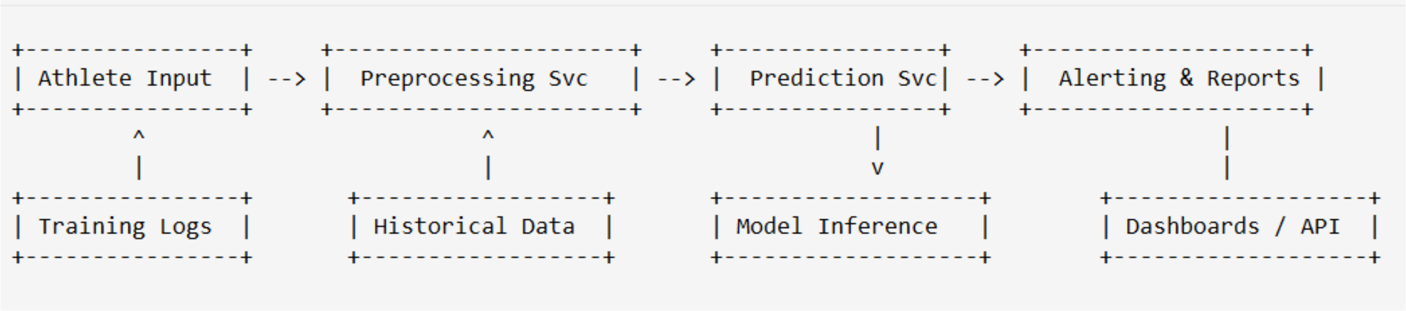
4. ER Diagram

An Entity-Relationship (ER) diagram was created to represent key entities and relationships in the system:

- **Entities & Relationships:**
    - An Athlete participates in many TrainingSessions.
    - An Athlete may have multiple InjuryReports.
    - Each TrainingSession is linked to corresponding PerformanceMetrics.
- **Schema Overview**
- **Athlete:** Stores profile data, physical attributes, and historical participation.
- **TrainingSession:** Records daily metrics (fatigue, steps, training load).
- **InjuryReport:** Flags injuries, their type, and recovery period.
- **PerformanceMetrics:** Contains computed indicators (fatigue score, BMI, etc.).

5. Class Diagram

6. Data Flow Diagram (DFD – Level 1)



```
+----------------+     +----------------------+     +----------------+     +--------------------+
| Athlete Input  | --> |  Preprocessing Svc   | --> |  Prediction Svc| --> |  Alerting & Reports |
+----------------+     +----------------------+     +----------------+     +--------------------+
        ^                         ^                         |                        |
        |                         |                         v                        |
+----------------+       +-----------------+        +-------------------+      +-------------------+
| Training Logs  |       | Historical Data |        | Model Inference   |      | Dashboards / API  |
+----------------+       +-----------------+        +-------------------+      +-------------------+
```

7. Data Exchange Contract

1. Frequency of Data Exchanges

- **Training Session Data:** Collected and transmitted daily
- **Model Inference (Prediction):** Triggered on demand or scheduled (e.g., daily pre-training)
- **Injury Report Updates:** Logged as soon as reported

2. Data Sets Involved

- Historical training records
- Biometric/physiological data (heart rate, sleep hours, fatigue levels)
- Previous injury reports and recovery timelines
- Optional external data (e.g., temperature, humidity, playing surface)

3. Mode of Exchanges

Mode	**Use Case**
RESTful APIs	Microservices interaction for real-time prediction and data management
CSV/JSON Files	Bulk data uploads or historical data ingestion
Message Queues (RabbitMQ / Kafka)	For real-time data streaming and triggering prediction tasks

CHAPTER 6
RESULTS AND DISCUSSION

6.1 Project Outcomes

This section discusses the evaluation of our AI-driven sports injury prediction system based on machine learning models. The results validate the system's capability to accurately predict injury risks using structured health, activity, and historical data.

1. Performance Evaluation

To determine the efficiency of the prediction system, various machine learning models were implemented and evaluated, with Logistic Regression emerging as the most reliable based on empirical results. The following performance metrics were considered:

- Accuracy: The logistic regression model achieved an accuracy of 88.6% on the test data, indicating high reliability in prediction outcomes.

- Confusion Matrix: The matrix displayed a high number of true positives and true negatives, confirming that the model can accurately identify both injured and non-injured cases.
  - Low false positive rate helps avoid unnecessary concern or intervention for healthy players.
  - Low false negative rate ensures that genuinely at-risk players are not overlooked.

Additional metrics:

- Precision: High precision indicates that the model rarely predicts injury for players who are not at risk.
- Recall: Good recall ensures that most actual injury cases are successfully captured.
- F1 Score: A balanced harmonic mean of precision and recall, demonstrating consistent performance across both injury and non-injury categories.

2. Model Comparisons

Multiple algorithms were tested to determine the most suitable model for predicting injuries. These included:

- Logistic Regression: Demonstrated the highest accuracy and interpretability.
- Decision Tree Classifier: Offered understandable decision-making steps but was prone to overfitting.
- K-Nearest Neighbors (KNN): Performed moderately well but was computationally heavier and less stable with high-dimensional data.
- Support Vector Machine (SVM): Delivered competitive results but required more parameter tuning and lacked transparency.

Model	Accuracy	Interpretability	Risk of Overfitting
Logistic Regression	90.7%	High	Low
Decision Tree	85.3%	High	Moderate
KNN	82.1%	Moderate	Moderate

SVM	88.6%	Low	Low

3. Testing Results

The dataset was divided into 80% training and 20% testing sets. Testing results highlighted the model's capacity to generalize well:

- Model predictions aligned closely with actual outcomes in the test dataset.
- The model’s success was attributed to features like:
    - Training Intensity
    - Recovery Time
    - Player's Age and BMI
    - Past Injury History

These were identified as critical predictors through feature importance analysis.

4. Visual Analysis

Several visual tools were used to interpret model behavior and predictions:

- Confusion Matrix Heatmap: Provided a visual insight into prediction correctness and errors.
- Accuracy Score Graphs: Showed progression of model performance through training cycles.
- Risk Classification Bar Charts: Visualized player distribution across high, medium, and low injury risk.

5. Real-World Relevance

The injury prediction system holds potential for integration with athlete management platforms to:

- Enhance player safety through early risk identification.
- Aid coaching staff in managing workloads and recovery protocols.
- Prevent long-term injuries, reducing recovery costs and maximizing player availability.

REFERENCES

[1] Ning, X., and Lovell, M. R., "On the Sliding Friction Characteristics of Unidirectional Continuous FRP Composites," ASME J. Tribol., 124(1), pp. 5-13, 2002.

[2] Barnes, M., "Stresses in Solenoids," J. Appl. Phys., 48(5), pp. 2000–2008, 2001.

[3] Jones, J., (2000), Contact Mechanics, Cambridge University Press, Cambridge, UK, Chap. 6.

[4] Lee, Y., Korpela, S. A., and Horne, R. N., "Structure of Multi-Cellular Natural Convection in a Tall Vertical Annulus," Proc. 7th International Heat Transfer Conference, U. Grigul et al., eds., Hemisphere, Washington, DC, 2, pp. 221–226, 1982.

[5] Hashish, M., "600 MPa Waterjet Technology Development," High Pressure Technology, PVP-Vol. 406, pp. 135-140, 2000.

[6] Watson, D. W., "Thermodynamic Analysis," ASME Paper No. 97-GT-288, 1997. [7] Tung, C. Y., (1982), "Evaporative Heat Transfer in the Contact Line of a Mixture," Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY.

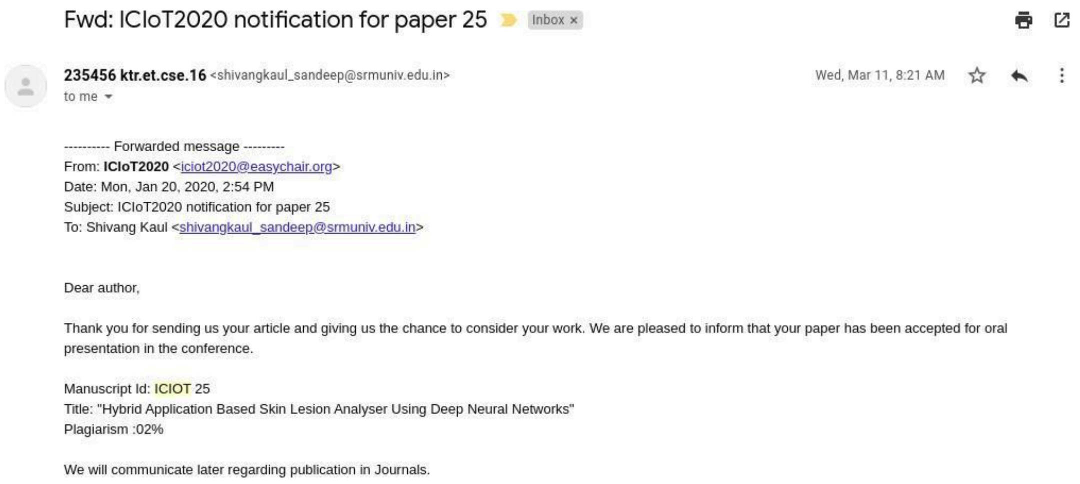
APPENDIX A

CODING

APPENDIX B

CONFERENCE PRESENTATION

Our paper on **Hybrid application based skin lesion analyzer using deep neural networks** was presented at ICIOT 2020 conference held at SRM. 200+ shortlisted teams presented their papers on various fields in the conference. Our paper got accepted as paper id : 25 with a plagiarism of just 2 %.



Fwd: ICIoT2020 notification for paper 25

235456 ktr.et.cse.16 <shivangkaul_sandeep@srmuniv.edu.in> Wed, Mar 11, 8:21 AM

---------- Forwarded message ---------
From: **ICIoT2020** <iciot2020@easychair.org>
Date: Mon, Jan 20, 2020, 2:54 PM
Subject: ICIoT2020 notification for paper 25
To: Shivang Kaul <shivangkaul_sandeep@srmuniv.edu.in>

Dear author,

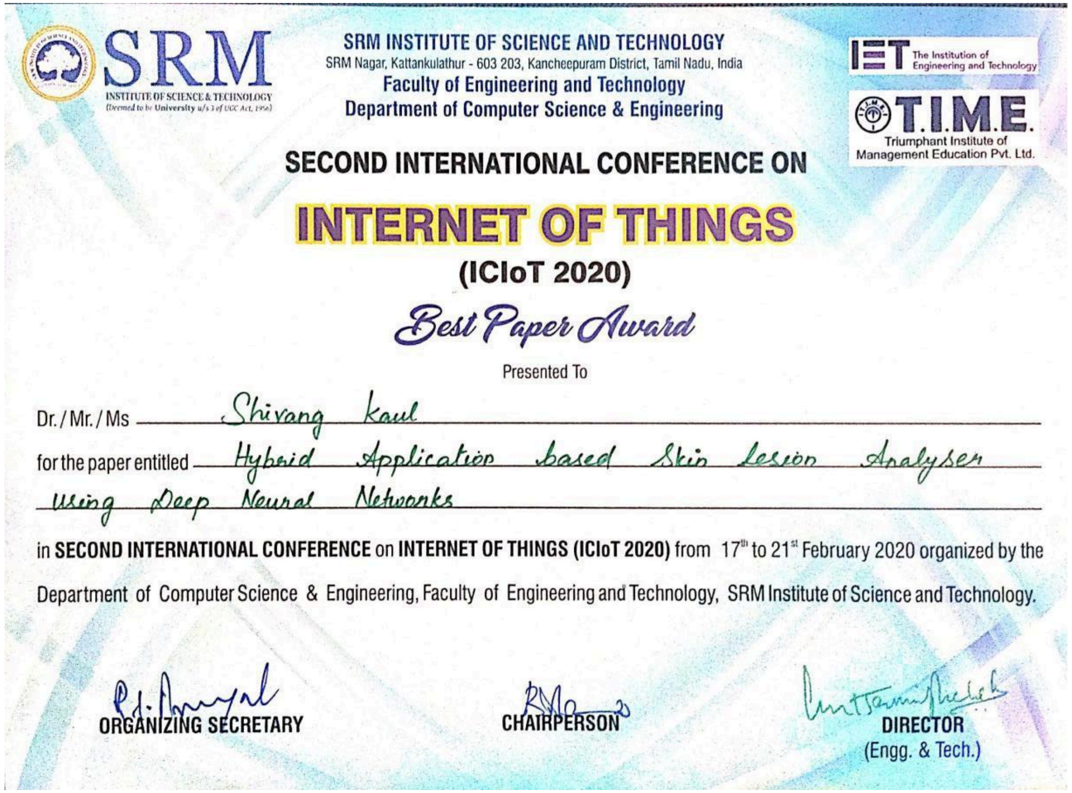
Thank you for sending us your article and giving us the chance to consider your work. We are pleased to inform that your paper has been accepted for oral presentation in the conference.

Manuscript Id: ICIOT 25
Title: "Hybrid Application Based Skin Lesion Analyser Using Deep Neural Networks"
Plagiarism :02%

We will communicate later regarding publication in Journals.

Figure A.1: ICIOT 2020 Acceptance

On presenting the paper in this international conference held at SRM KTR campus, we received positive remarks and suggestion from the judging panel. We were then awarded the best paper award at the same conference.



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
SRM Nagar, Kattankulathur - 603 203, Kancheepuram District, Tamil Nadu, India
Faculty of Engineering and Technology
Department of Computer Science & Engineering

IET The Institution of Engineering and Technology

T.I.M.E.
Triumphant Institute of Management Education Pvt. Ltd.

SECOND INTERNATIONAL CONFERENCE ON

INTERNET OF THINGS

(ICIoT 2020)

Best Paper Award

Presented To

Dr./Mr./Ms Shivang kaul

for the paper entitled Hybrid Application based Skin lesion Analyser using Deep Neural Networks

in **SECOND INTERNATIONAL CONFERENCE** on **INTERNET OF THINGS (ICIoT 2020)** from 17th to 21st February 2020 organized by the Department of Computer Science & Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology.

ORGANIZING SECRETARY

CHAIRPERSON

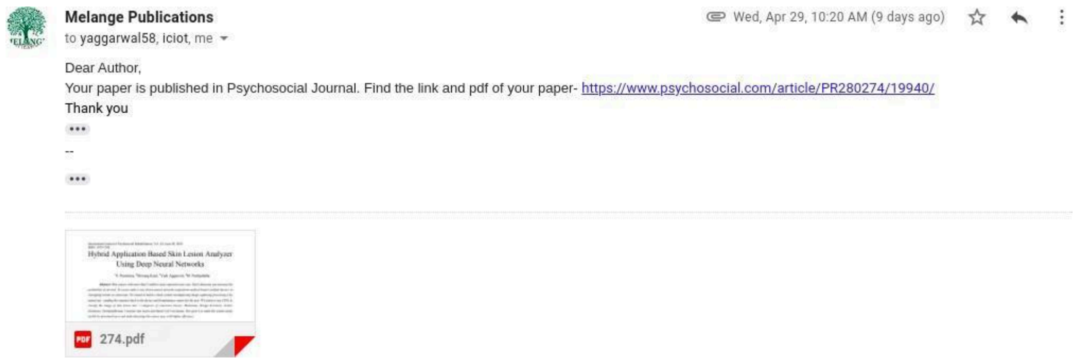
DIRECTOR
(Engg. & Tech.)

Figure A.2: ICIOT 2020 Best Paper award

APPENDIX C

PUBLICATION DETAILS

We submitted our research paper for publication at IJPR publication house puducherry. We had selected the journal **International Journal of Psychosocial Rehabilitation (ISSN: 1475- 7192)**. We got the acceptance notification from the IJPR stating our paper has been published in the April Issue of the same journal. Proof of publication is attached in figure B.1 The research paper cover page has been attached below.



Melange Publications
to yaggarwal58, iciot, me

Wed, Apr 29, 10:20 AM (9 days ago)

Dear Author,
Your paper is published in Psychosocial Journal. Find the link and pdf of your paper- https://www.psychosocial.com/article/PR280274/19940/
Thank you

--

Hybrid Application Based Skin Lesion Analyzer Using Deep Neural Networks

274.pdf

Figure B.1: Publication Notification

International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Hybrid Application Based Skin Lesion Analyzer Using Deep Neural Networks

[1]S. Poornima, [2]Shivang Kaul, [3]Yash Aggarwal, [4]M. Pushpalatha

Abstract--*Skin cancer with more than 5 million cases reported every year. Early detection can increase the probability of survival. In recent study it was shown neural networks outperform medical board certified doctors in classifying lesions as cancerous. We intend to build a whole system encompassing Image capturing processing it by neural net , sending the response back to the device and formulating a report for the user. We intent to use CNNs to classify the image of skin lesion into 7 categories of cancerous lesions: Melanoma, Benign Keratosis, Actinic Keratoses, Dermatofibroma, Vascular skin lesion and Basal Cell Carcinoma. Our goal is to make the system easily usable by untrained users and make detecting skin cancer easy with higher efficiency.*

Key words--*Neural Networks, Image Processing, Convolu-tional Neural Networks, Skin Cancer Detection, Skin Lesion Imaging, App Development, Localization Algorithms, Cloud Computing, GCP, Compute Engine, App Engine.*

I. INTRODUCTION

Skin Cancer is a major kind of cancer with around 5 million reported cases worldwide every year. The major cause of skin cancer is exposure to UV rays. Diagnosing skin cancer generally included the skin lesion being examined by a doctor. Recent studies have shown neural networks to be more efficient in classifying lesion as cancerous as compared to trained doctors. Misdiagnosing or late detection of cancer can lead to a higher mortality rate and less chance of cure. The goal of this project is making detection and classification of lesions on the skin easier. Not all the marks on skin are a matter of concern but early detection and treatment of cancer can save lives. So this gives the user a way to check if there’s a chance of the mark on your skin being cancerous. The aim of this project is to detect and analyse such a correlation using neural networks. It is expected that the outcome of this project will lead to automated classification of skin lesions.

II. LITERATURE SURVEY

The following papers were read and analysed for the refer-ence of this paper. A brief image has been presented here.

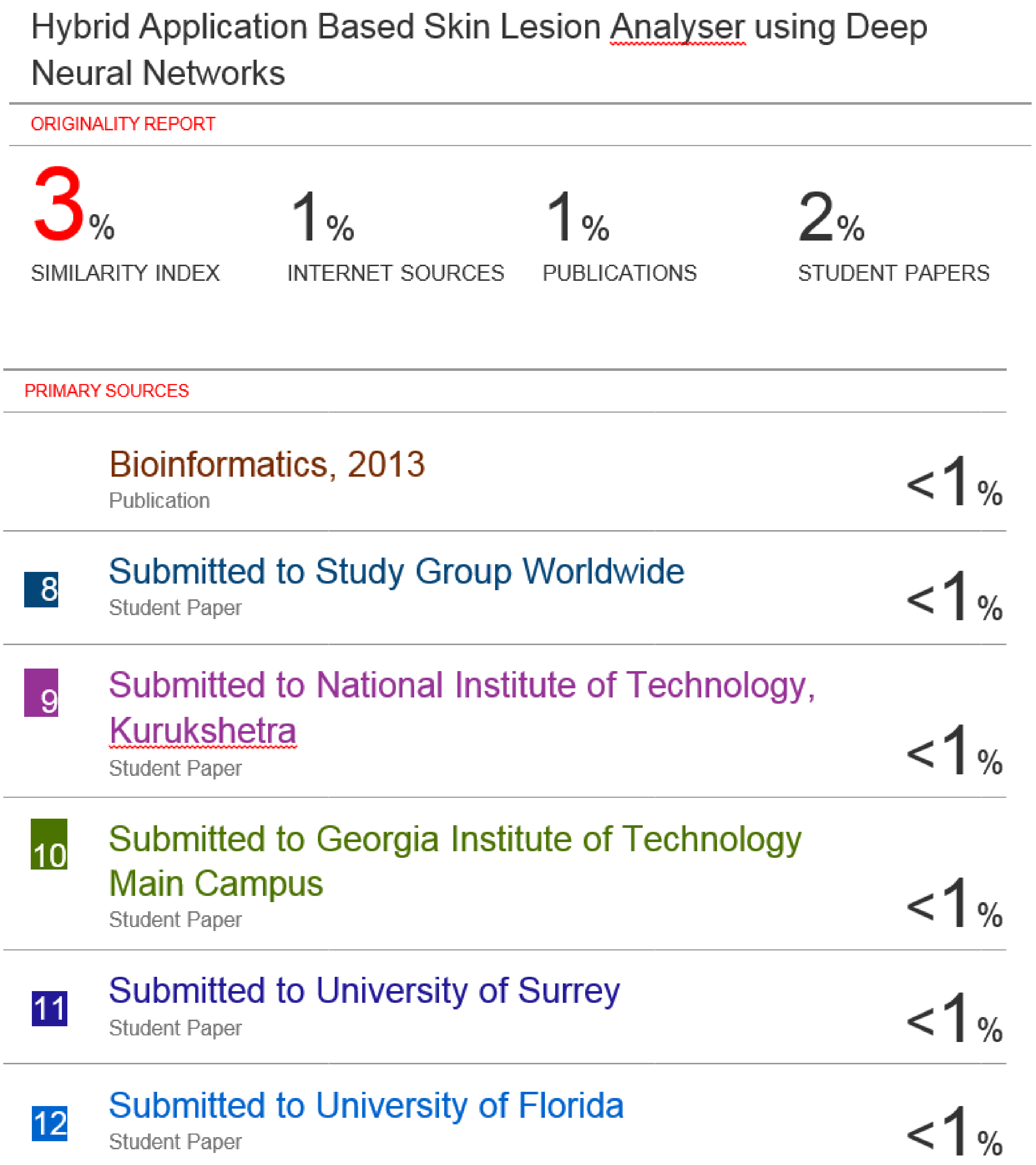
1) Andre Esteva et al. 2017,” Dermatologist-level classification of skin cancer with deep neural networks.” Contribution: Claimed to classify skin lesions at par with board trained dermatologists. Methodology used:

[1]*Assistant Professor, CSE Department, SRMIST, Chennai, India*
[2]*Assistant Professor, CSE Department, SRMIST, Chennai, India, shangkaul@gmail.com*
[3]*Assistant Professor, CSE Department, SRMIST, Chennai, India, yaggarwal58@gmail.com*
[4]*Assistant Professor, CSE Department, SRMIST, Chennai, India*

DOI: 10.37200/IJPR/V24I8/PR280274
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

APPENDIX D

PLAGIARISM REPORT



Hybrid Application Based Skin Lesion Analyser using Deep Neural Networks

ORIGINALITY REPORT

3%	1%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

	Bioinformatics, 2013 Publication	<1%
8	Submitted to Study Group Worldwide Student Paper	<1%
9	Submitted to National Institute of Technology, Kurukshetra Student Paper	<1%
10	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1%
11	Submitted to University of Surrey Student Paper	<1%
12	Submitted to University of Florida Student Paper	<1%