Department of Electrical and Computer Engineering
Pak-Austria Fachhochschule: Institute of Applied Sciences & Technology
School of Computing Sciences

COMP-240: Machine Learning

Complex Computing Problem (CCP)

Course Title: Machine Learning	Course Code: COMP-240
Batch:	Credit Hours: 2 CHs
Instructor: Dr. Abid Ali	Weightage: 5%
Semester: 5th (BS AI)	Time
Date: 22-11-2025	CLO-PLO: 2,3-3,4

CCP Title: A Comparative Machine Learning Framework for Medical Diagnostic Prediction

1. CCP Statement:

1. CCP Statement:

The primary goal of this CCP is to develop and critically evaluate a predictive diagnostic system for a significant healthcare challenge, such as breast cancer, heart disease, or diabetes. The aim is to move beyond a single-model approach by implementing a comprehensive framework that leverages multiple core machine learning algorithms to diagnose diseases from patient clinical data. This project underscores the critical importance of model selection, interpretability, and accuracy in medical diagnostics, where a misprediction can have serious consequences. The challenge lies in handling real-world clinical data, which is often imbalanced, contains missing values, and has features of varying scales and relevance. Students must justify their data preprocessing, feature engineering, and hyperparameter tuning decisions through rigorous experimentation and validation using publicly available medical datasets (e.g., from UCI ML Repository, Kaggle). Failure to adequately compare and validate models may lead to over-reliance on a suboptimal algorithm, resulting in poor diagnostic accuracy, lack of clinician trust, and potential risks to patient safety. Therefore, it is imperative to provide a data-driven justification for the best-performing model in the given context.

2. Range of Problem Solving for this CCP activity is mapped on the following attributes:

Sr. No.	Characteristic	A Complex Computing Problem is a computing problem having some or all the following characteristics:	Statements (CCP Machine Learning)	Mapping
WP1	Range of conflicting requirements	Involves wide-ranging or conflicting technical, computing, and other issues	This CCP involves conflicting requirements between model complexity and interpretability (e.g., a highly accurate ANN vs. an interpretable Decision Tree), computational cost vs. performance, and handling class imbalance without distorting the data's real-world distribution.	✓
WP2	Depth of analysis required	Has no obvious solution, and requires conceptual thinking and innovative analysis to formulate suitable abstract models	There is no single "best" algorithm for all medical datasets. Students must conceptually analyze the problem, hypothesize which models (e.g., SVM for high-dimensional spaces, Random Forest for robustness) might perform best, and	✓

			innovate in their evaluation strategy.	
WP3	Depth of knowledge required	A solution requires the use of in-depth computing or domain knowledge and an analytical approach that is based on well-founded principles	The solution requires in-depth knowledge of multiple ML algorithms (DT, RF, SVM, NB, ANN, KNN), their theoretical foundations, metrics for imbalanced data (Precision, Recall, F1-Score, AUC-ROC), and an understanding of the medical domain to interpret features.	✓
WP4	Familiarity of issues	Involves infrequently encountered issues	Students will deal with the "black-box" nature of models like ANN and RF, the curse of dimensionality for KNN, and the challenge of making a Naive Bayes model perform well on data that violates its core independence assumption.	✓
WP5	Level of problem	Is outside problems encompassed by standards and standard practice for professional computing	Standard practice is to apply one or two models. This project requires building a multi-model, end-to-end comparative analysis pipeline, which is a non-standard, research-oriented task.	✓

WP8	Interdependence	Is a high-level problem possibly including many component parts or subproblems	The CCP consists of many interdependent components: data loading, preprocessing, feature scaling, model implementation, hyperparameter tuning, model evaluation, and result interpretation, where the output of one stage is critical for the next.	✓

3. Deliverables:

- **Dataset:** A publicly available medical diagnostic dataset (e.g., Wisconsin Breast Cancer, Cleveland Heart Disease, Pima Indians Diabetes).
- **Python Implementation:** A complete, well-documented codebase that includes:
  - Data loading and exploratory data analysis (EDA).
  - Data preprocessing (handling missing values, encoding, feature scaling).
  - Implementation and training of at least **four** of the following: Decision Tree, Random Forest, SVM, Naive Bayes, ANN, KNN.
  - Hyperparameter tuning for at least two models.
  - A comprehensive model evaluation and comparison module.
- **Visualization:** Comparative visualizations (e.g., ROC curves, confusion matrices, feature importance charts).
- **Detailed Report and Demonstration:** A report justifying every design decision and a live demonstration of the system.

Through the course of this CCP, students should be able to partially attain some or all of the following graduate attributes: **GA3: Problem Analysis, GA4: Design/Development of Solutions**

Course Learning Outcomes (CLOs):
At the end of the course the students will be able to:

1. **Apply** data preprocessing and feature engineering techniques to prepare real-world medical data for machine learning models. (PLO-3, BT Level 2)
2. **Analyze** the underlying assumptions and mechanics of various machine learning algorithms to determine their suitability for a classification task. (PLO-4, BT Level 4)

3. **Design and Develop** a robust machine learning pipeline that implements, trains, and evaluates multiple models to solve a diagnostic prediction problem. (PLO-5, BT Level 6)

4. Project Description:

In this project, students will design and implement a comparative machine learning framework to predict the presence or absence of a disease. The ability to accurately and reliably diagnose conditions using automated systems is a cornerstone of modern healthcare informatics.

Students will select a medical dataset, preprocess it to handle real-world data issues, and implement a suite of machine learning algorithms. The core of the project is a rigorous, empirical comparison to determine which model offers the best trade-off between accuracy, precision, and interpretability for the specific dataset.

Students will gain:

- Practical experience in building an end-to-end ML pipeline for a critical application.
- Deep understanding of the strengths and weaknesses of different ML algorithms.
- Insight into handling imbalanced datasets common in medical diagnostics.
- Skills in model evaluation, hyperparameter tuning, and result interpretation.

Key Tasks:

- **Research and Dataset Selection:** Research and select an appropriate medical dataset. Understand its features and the clinical context.
- **Data Preprocessing & EDA:** Clean the data, handle missing values, encode categorical variables, and perform feature scaling. Conduct EDA to understand data distributions and correlations.
- **Model Implementation:** Implement at least four ML models from the listed algorithms.
- **Model Training & Tuning:** Split the data, train the models, and perform hyperparameter tuning (e.g., using GridSearchCV or RandomizedSearchCV) for at least two models to optimize performance.
- **Evaluation & Comparison:** Evaluate all models using a suite of metrics (Accuracy, Precision, Recall, F1-Score, AUC-ROC). Create visual comparisons.
- **Documentation and Presentation:** Document the entire process, justify model selection and parameter choices, and present findings, concluding with a recommendation for the best model for the task.

Submission Deadlines:

- **Project Submission Deadline:** 10 November 2025

- **Presentation & Demo:** 05 and 10 December 2025

5. Sustainable Development Goals Mapping of CCP

Sr	SDGs	Mapping
8	Decent Work and Economic Growth	X
9	Industry, Innovation and Infrastructure	X
17	Partnerships for the Goals	X

Rubrics for Assessment of the CCP

I. Implementation (0–25)

Grade	Score	Description
Unacceptable	0–5	Minimal or non-functional code. Only one model implemented.
Just Acceptable	5–10	Code implements 2 models with basic preprocessing. Major errors present.
Basic	10–15	Functional pipeline for 3-4 models with standard preprocessing. Limited hyperparameter tuning.
Good	15–20	Full, clean pipeline for 4+ models. Proper train/test split, feature scaling, and hyperparameter tuning for at least 2 models.
Excellent	20–25	Robust, well-documented pipeline. Advanced techniques (e.g., handling class imbalance, feature selection). Exceeds the minimum model requirement with insightful implementation.

II. Demonstration (0–25)

Grade	Score	Description
Unacceptable	0–5	Cannot run the system or explain the code.
Just Acceptable	5–10	Demonstrates system with major guidance; explanations are unclear or incorrect.
Basic	10–15	Can run the system and provide a basic, mostly correct explanation of the workflow.

Good	15–20	Clear demonstration; explains the code and results correctly for all major components.
Excellent	20–25	Flawless demonstration; provides insightful commentary on why certain models performed better/worse, linking it to theory.

III. Presentation (0–25)

Grade	Score	Description
Unacceptable	0–5	No presentation or completely incoherent.
Just Acceptable	5–10	Slides are incomplete; presentation is disorganized and unclear.
Basic	10–15	Covers most key points; explanation is understandable but lacks depth.
Good	15–20	Clear, well-structured slides and presentation. Correctly interprets results and justifies decisions.
Excellent	20–25	Professional, engaging, and polished presentation. Provides deep insights and a compelling argument for the final model recommendation.

IV. Project Report (0–25)

Grade	Score	Description
Unacceptable	0–5	No report or extremely incomplete.
Just Acceptable	5–10	Report is missing major sections (e.g., methodology, results). Poor structure and numerous errors.
Basic	10–15	Report includes all sections but lacks depth in analysis and justification.
Good	15–20	Complete, well-written report with minor mistakes. Clearly explains the process and results.
Excellent	20–25	High-quality report that is detailed, well-analyzed, and professionally presented. Includes critical discussion on model comparison and limitations.