Priority	Project Code	Thoughts
1	TB05: Time-series classification	Time series Classification is a Machine Learning problem where the goal is to predict categorical labels for the time series sequences. This entails learning the underlying patterns and relationships within these sequences in order to accurately classify them. There are various approaches from *feature-based methods* using decision trees, *deep learning techniques* using CNN, RNN, LSTM & *distance-based methods* which use similarity metrics like Dynamic Time Warping to classify the time series. Time series is a category of data which spans across most domains from finance, healthcare, environmental sciences to many others. The use of such versatile data can be broken down into– forecasting & classification. Classification of time series enhances forecasting capabilities, anomaly detection and producing data-driven insights. The Problem statement involves applying KNN with DTW on the Human Activity Recognition dataset (HAR), aimed to classify sensor readings into human activity categories like sitting, walking, lying. This approach faces computational challenges with roughly 16.3 billion calculations for only 500 observations for training and testing [1]. Proposed solutions include application of KNN-IS which is an iterative SPARK-based design for KNN classifier [2], and constraints to reduce DTW calculations. This combination is effective to handle non-linear relationships and temporal distortions. KNN performs localized decision making and DTW captures temporal similarities. KNN is chosen over other clustering algorithms like DBSCAN because it works better for non-spherical data and operates on distances between data points unlike DBSCAN which relies on the data density. Additionally, exploring Fuzzy K-Nearest Neighbours clustering algorithm which outperforms KNN algorithm for overlapping categories & reduces risk of misclassifications [3]. Potentially integrating DTW as a distance measure. We wish to explore other datasets as

		well to provide a comprehensive analysis [5]. References: [1]https://github.com/markdregan/K-Nearest-Neighbors-with-Dynamic-Time-Warping/tree/master [2] https://github.com/JMailloH/kNN_IS [3] A Fuzzy K-nearest neighbour algorithm [4] Human Activity Recognition using Smartphone [5] Other possible datasets
2	PB01: Predicting Contact Maps in Bioinformatics	Our team is particularly interested in using contact map predictions for DNA and molecular computing. This area drives our engagement with the problem, as it's essential for designing computational elements from biological materials like DNA origami or DNA-based logic gates. These innovations aim to revolutionize computing by merging biological understanding with computational power. Additionally, contact map predictions play a crucial role in drug design, enabling the identification of interaction sites to accelerate therapeutic development. Similarly, they offer insights into proteins associated with diseases like cancer and neurodegenerative disorders, facilitating targeted interventions by understanding molecular mechanisms. Initial Approach to Solving the Problem: ● Data Preprocessing: Our strategy includes a thorough investigation of dimensionality and feature reduction techniques, such as UMAP, MDS, ISOMAP, LASSO feature selection, and the chi-square test. We aim to not only utilize these techniques individually but also investigate their combinations and the sequence in which they are applied.[3] ● Addressing Imbalance: To enhance the solution, this project will utilize Apache

Spark for large dataset handling and compare methods such as SMOTE and ADASYN for addressing imbalanced datasets with their ability to create synthetic data [4].

- Algorithm Selection and Optimization: We will explore boosted trees applying the focal loss function to address class imbalances and exploring Cost-Sensitive Learning for minority class sensitivity. Additionally, ensemble methods including XGBoost and GBM will be compared with SVM, given their superior performance attributes such as gradient boosting, regularization, handling of complex relationships, weighted sampling, tuning flexibility, and scalability for parallel computing. The potential of kernel SVM with adjusted class weights will also be evaluated [2].

Through comparing and contrasting various combinations of these proposed models, the aim is to improve performance.

References:

1. Isaac Triguero, Sara del Río, Victoria López, Jaume Bacardit, José M. Benítez, Francisco Herrera,ROSEFW-RF: The winner algorithm for the ECBDL’14 big data competition: An extremely imbalanced big data bioinformatics problem,Knowledge-Based Systems,Volume 87,2015,Pages 69-79,ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2015.05.027.

2. Li B, Zhang N, Wang Y-G, George AW, Reverter A and Li Y (2018) Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. Front. Genet. 9:237. https://doi.org/10.3389/fgene.2018.00237

3. Front. Bioinform., 27 June 2022
Sec. Integrative Bioinformatics
Volume 2 - 2022 | https://doi.org/10.3389/fbinf.2022.927312

4. Priyadharshini, M.; Banu, A.F.; Sharma, B.; Chowdhury, S.; Rabie, K.; Shongwe, T. Hybrid

		Multi-Label Classification Model for Medical Applications Based on Adaptive Synthetic Data and Ensemble Learning. Sensors 2023, 23, 6836. https://doi.org/10.3390/s23156836
3	PB03: A Classification Model for Healthcare Sentiments on Twitter	This project is interesting because it aims to leverage Twitter data to extract health-related tweets and categorize them into relevant health keywords using big data frameworks. Twitter provides real-time insights into social events, emerging threats, and epidemics, which can inform decision-making in health planning and execution, thus optimizing program success while reducing costs and time. The dataset, initially unstructured and text-based, requires extensive natural language processing (NLP) for pre-processing and labeling to enable comparison of machine learning (ML) models. Data cleaning procedures, including special character removal, stop-word elimination, lowercase conversion, and spell-checking, will be employed to improve data quality. To address the time and cost constraints of human labeling, this project investigates ML algorithms approach. Techniques from seminal research papers are explored, including MetaMap for medical term extraction in UMLS Meta-thesaurus and Expectation Maximization (EM) clustering for word clustering (Detailed in Ref : 2). Additionally, NLP and Word Embedding Based Clustering Classification methods, utilizing word2vec and similarity measures, are considered (Detailed in Ref : 3).For Classification, various ML algorithms, such as Naïve Bayes, Random Forest, and Support Vector Classification (SVC), will be evaluated, due to their performance in handling large datasets, imbalanced data, faster learning and ability to run in CPUs (compared to ANN), and interpretability (Detailed in Ref : 5,6). A study by Manias et al. (2023) discusses transformers-based pre-trained models like RoBERTa and TwHIN-BERT for Twitter dataset classification, focusing on zero-shot classification. Pre-trained LLMs are considered due to their exceptional performance in the NLP world. However, the application and time complexity in big data setups warrant further exploration (Detailed in Ref : 4).Model performance will be

		assessed using balanced accuracy and overall accuracy metrics due to potential dataset imbalance. References: (1) Dataset -https://archive.ics.uci.edu/ml/datasets/Health%2BNews%2Bin%2BTwitter (2) Lu, Y., Zhang, P., Liu, J., Li, J., & Deng, S. (2013). Health-Related Hot Topic Detection in Online Communities Using Text Clustering. PLOS ONE, 8(2), e56221. https://doi.org/10.1371/journal.pone.0056221 (3) X. Dai, M. Bikdash and B. Meyer, "From social media to public health surveillance: Word embedding based clustering method for twitter classification," SoutheastCon 2017, Concord, NC, USA, 2017, pp. 1-7, doi: 10.1109/SECON.2017.7925400. (4) Manias, G., Mavrogiorgou, A., Kiourtis, A. et al. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Comput & Applic 35, 21415–21431 (2023). https://doi.org/10.1007/s00521-023-08629-3 (5) Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, Paul Cotae, Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset, Expert Systems with Applications,Volume 212,2023,118715,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.118715. (6)https://www.sciencedirect.com/science/article/pii/S1532046414000628 (7)https://hcis-journal.springeropen.com/articles/10.1186/s13673-017-0116-3 (8)https://www.researchgate.net/publication/276456888_Twitter_sentiment_classification_for_mea

		suring_public_health_concerns