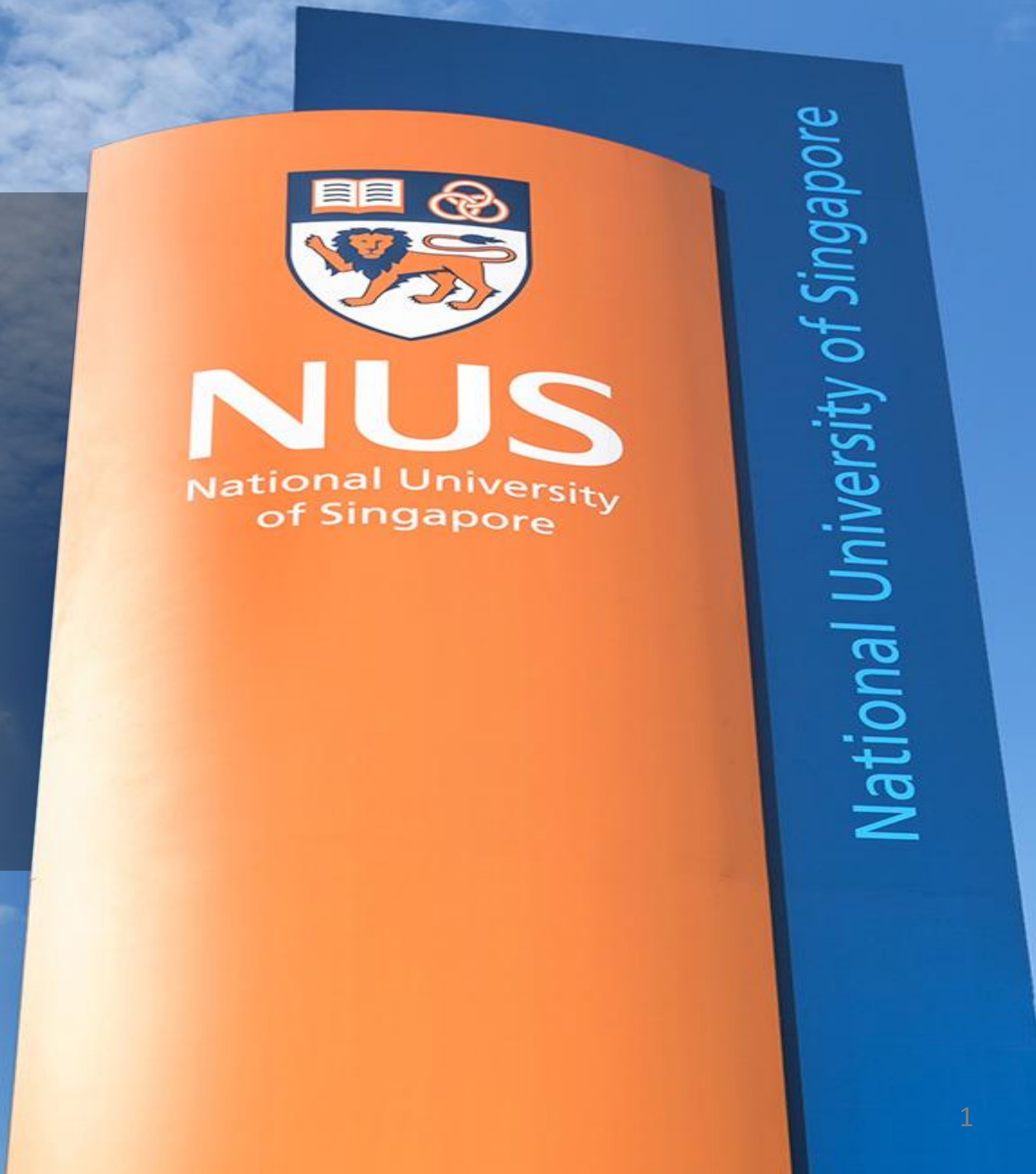
BS6213
Reflective scientist

Assessment arrangements for the part led by Professor Wong Limsoon


NUS
National University of Singapore
National University of Singapore

Course objective

Sharing with students how scientists do their thinking

The research / papers chosen by an instructor mainly serve as scaffolds to facilitate this sharing

Research papers typically outline the problem being addressed, the solution, and the results. They rarely delve into the thought processes behind the work. Our aim is to bridge this gap

Course plan

Details available on https://www.comp.nus.edu.sg/~wongls/courses/bs6213/2025/readme.htm

Session #1, 5 Jan 2026 and

Protein function prediction and some lessons for classifier performance evaluation

Session #2, 12 Jan 2026

Gene expression analysis and some lessons for statistical hypothesis testing

Session #3, 19 Jan 2026

The data science of PCA: Myths, misuses, and missed signals

Session #4, 26 Jan 2026

Insight + logic = elegant solutions

Assessment plan

3 homeworks for sessions 1, 2, 3

3 x 25% marks for the reports

Class interactions

1 x 4% marks for interactions in session 1

3 x 7% marks for presentations & interactions in sessions 2, 3, 4

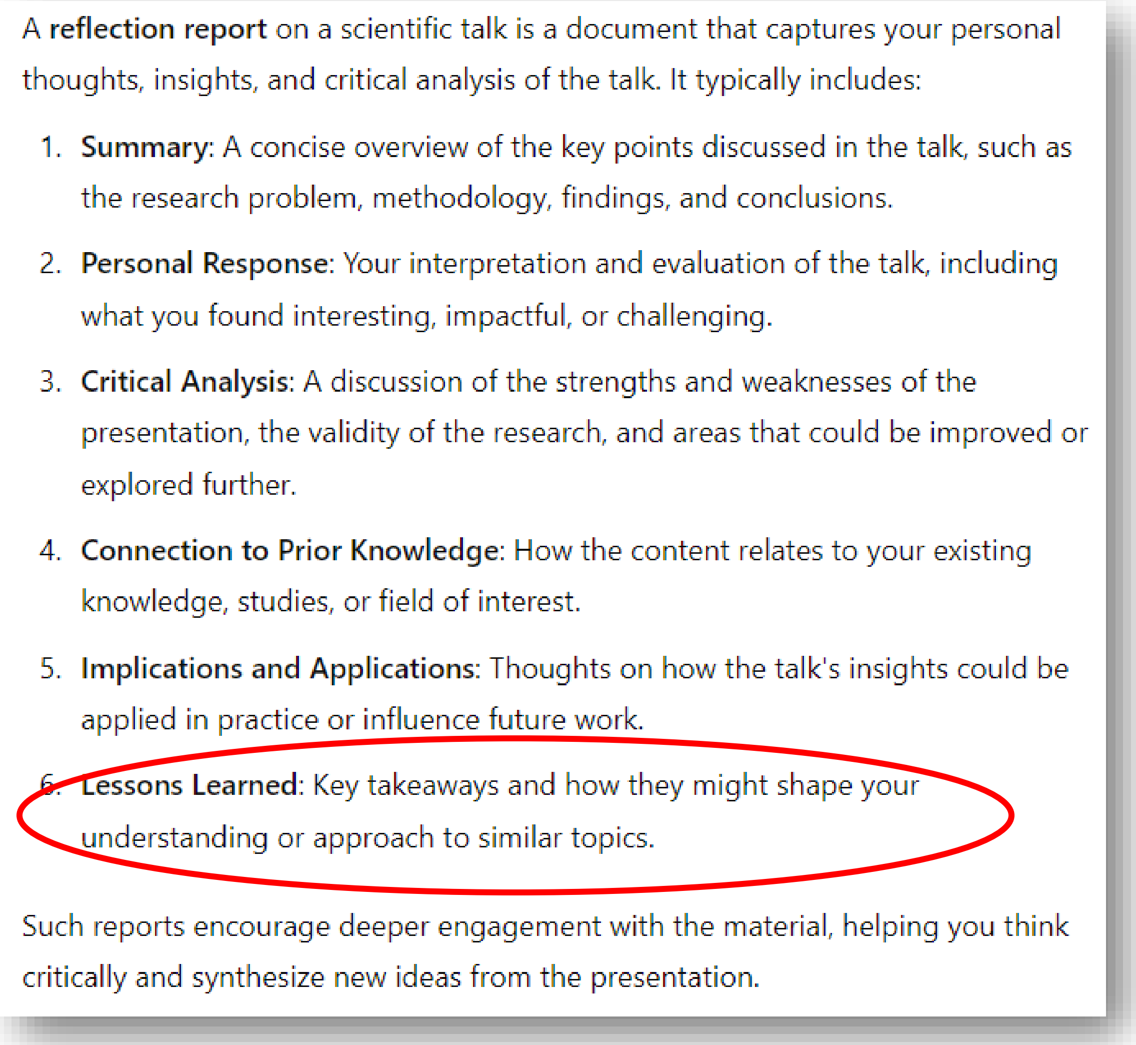
Please submit a reflection report (max 1 page) within 72 hours after each session to get interaction marks

Reflection report

1-page report submitted within 72 hours after a session

Provide your reflection on what you have learned from these interactions and from the session

Provide a record of the questions you have asked or responded to in a session



A **reflection report** on a scientific talk is a document that captures your personal thoughts, insights, and critical analysis of the talk. It typically includes:

1. **Summary**: A concise overview of the key points discussed in the talk, such as the research problem, methodology, findings, and conclusions.
2. **Personal Response**: Your interpretation and evaluation of the talk, including what you found interesting, impactful, or challenging.
3. **Critical Analysis**: A discussion of the strengths and weaknesses of the presentation, the validity of the research, and areas that could be improved or explored further.
4. **Connection to Prior Knowledge**: How the content relates to your existing knowledge, studies, or field of interest.
5. **Implications and Applications**: Thoughts on how the talk's insights could be applied in practice or influence future work.
6. **Lessons Learned**: Key takeaways and how they might shape your understanding or approach to similar topics.

Such reports encourage deeper engagement with the material, helping you think critically and synthesize new ideas from the presentation.

Homework #1, due 10/1/2026

submit by email to dcswls@nus.edu.sg

Read this paper, focusing on the way it evaluated the performance of the proposed classifier:

Yu et al., "Accurate prediction and key protein sequence feature identification of cyclins", Briefings in Functional Genomics, 22:411-419, 2023

Read the review report (attached below) produced by ChatGPT on the performance evaluation done by this paper.

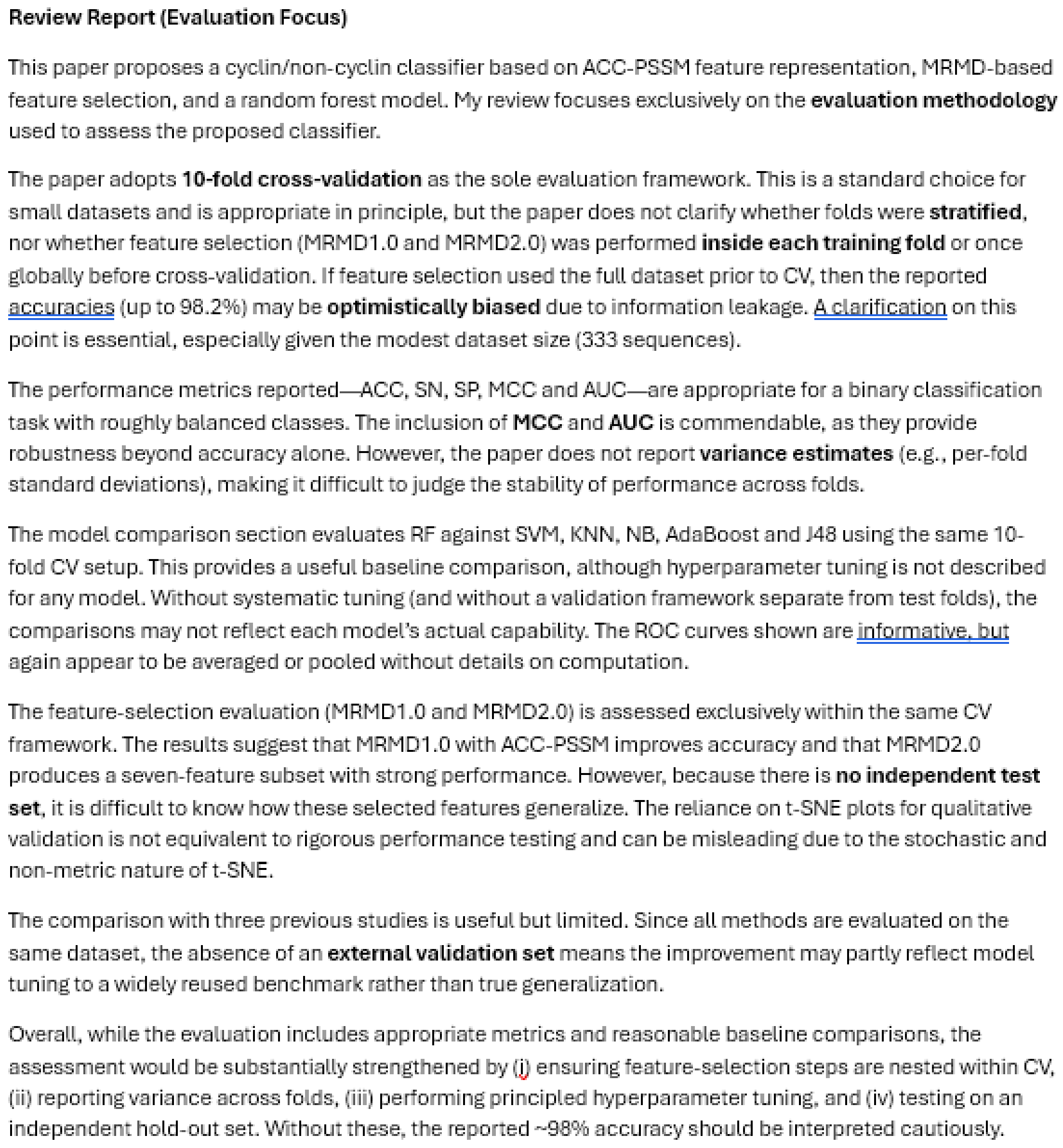
Answer the following questions:

1/ Which parts of the ChatGPT report, if any, do you agree with?

2/ Which parts of the ChatGPT report, if any, do you disagree with? Why?

3/ Are there any major flaws that you think should be pointed out but are missed by ChatGPT?

Make 5-minutes presentation to class on 12/1/2026



Review Report (Evaluation Focus)

This paper proposes a cyclin/non-cyclin classifier based on ACC-PSSM feature representation, MRMD-based feature selection, and a random forest model. My review focuses exclusively on the **evaluation methodology** used to assess the proposed classifier.

The paper adopts **10-fold cross-validation** as the sole evaluation framework. This is a standard choice for small datasets and is appropriate in principle, but the paper does not clarify whether folds were **stratified**, nor whether feature selection (MRMD1.0 and MRMD2.0) was performed **inside each training fold** or once globally before cross-validation. If feature selection used the full dataset prior to CV, then the reported accuracies (up to 98.2%) may be **optimistically biased** due to information leakage. A clarification on this point is essential, especially given the modest dataset size (333 sequences).

The performance metrics reported—ACC, SN, SP, MCC and AUC—are appropriate for a binary classification task with roughly balanced classes. The inclusion of **MCC** and **AUC** is commendable, as they provide robustness beyond accuracy alone. However, the paper does not report **variance estimates** (e.g., per-fold standard deviations), making it difficult to judge the stability of performance across folds.

The model comparison section evaluates RF against SVM, KNN, NB, AdaBoost and J48 using the same 10-fold CV setup. This provides a useful baseline comparison, although hyperparameter tuning is not described for any model. Without systematic tuning (and without a validation framework separate from test folds), the comparisons may not reflect each model's actual capability. The ROC curves shown are informative, but again appear to be averaged or pooled without details on computation.

The feature-selection evaluation (MRMD1.0 and MRMD2.0) is assessed exclusively within the same CV framework. The results suggest that MRMD1.0 with ACC-PSSM improves accuracy and that MRMD2.0 produces a seven-feature subset with strong performance. However, because there is **no independent test set**, it is difficult to know how these selected features generalize. The reliance on t-SNE plots for qualitative validation is not equivalent to rigorous performance testing and can be misleading due to the stochastic and non-metric nature of t-SNE.

The comparison with three previous studies is useful but limited. Since all methods are evaluated on the same dataset, the absence of an **external validation set** means the improvement may partly reflect model tuning to a widely reused benchmark rather than true generalization.

Overall, while the evaluation includes appropriate metrics and reasonable baseline comparisons, the assessment would be substantially strengthened by (i) ensuring feature-selection steps are nested within CV, (ii) reporting variance across folds, (iii) performing principled hyperparameter tuning, and (iv) testing on an independent hold-out set. Without these, the reported ~98% accuracy should be interpreted cautiously.

Homework #2
due 17/1/2026

submit by email to dcswls@nus.edu.sg

Read this paper, focusing on how it proposed to test synthetic-lethal gene pairs:

Sriganesh et al., "Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer", *Biology Direct*, 10:57, 2015

Read the review report (attached below) produced by ChatGPT discuss whether the proposed test is a good one.

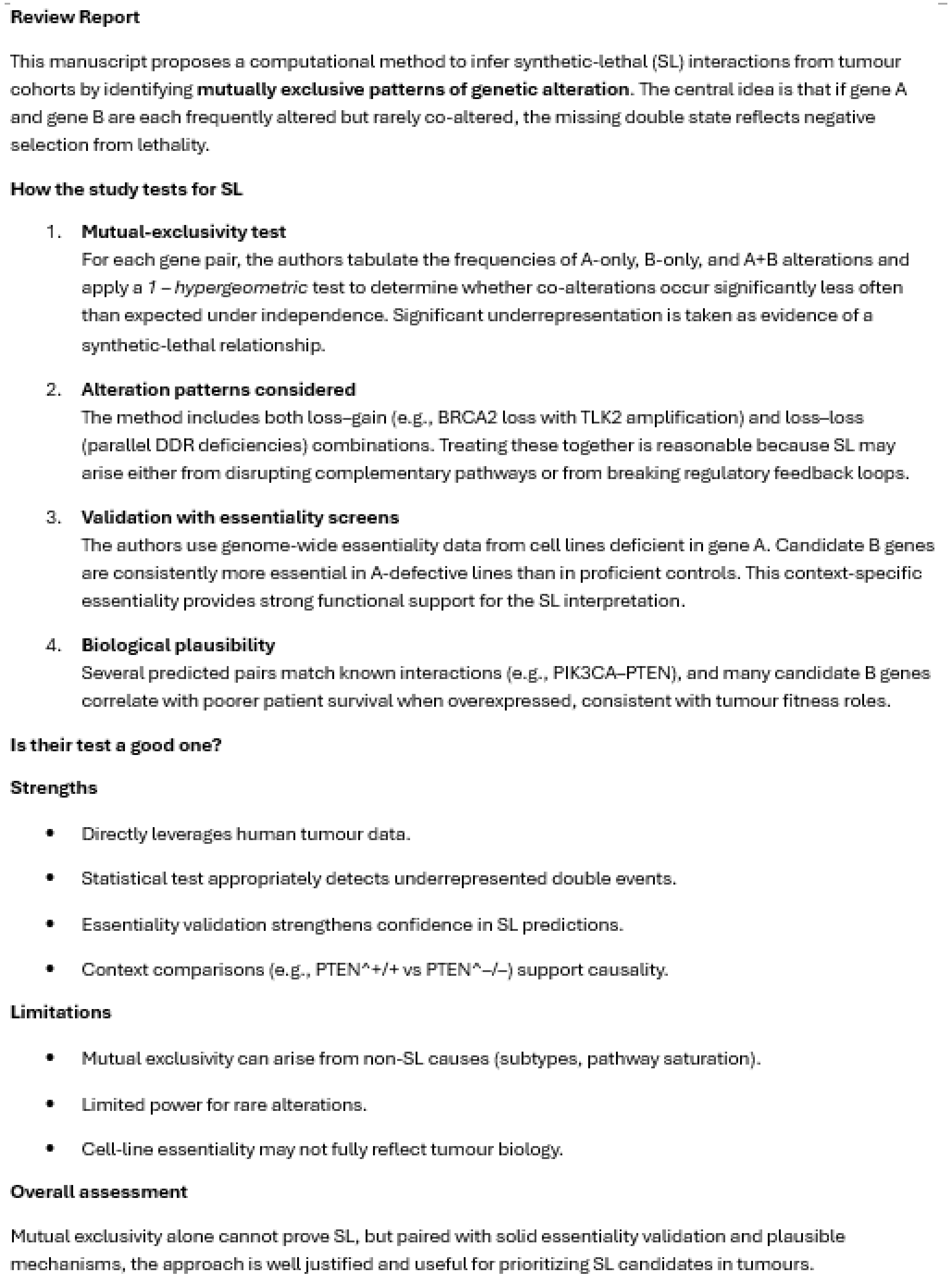
Answer the following questions:

1/ Which parts of the ChatGPT report, if any, do you agree with?

2/ Which parts of the ChatGPT report, if any, do you disagree with? Why?

3/ Are there any major flaws that you think should be pointed out but are missed by ChatGPT?

Make 5-minutes presentation to class on 19/1/2026



Review Report

This manuscript proposes a computational method to infer synthetic-lethal (SL) interactions from tumour cohorts by identifying **mutually exclusive patterns of genetic alteration**. The central idea is that if gene A and gene B are each frequently altered but rarely co-altered, the missing double state reflects negative selection from lethality.

How the study tests for SL

1. **Mutual-exclusivity test**
For each gene pair, the authors tabulate the frequencies of A-only, B-only, and A+B alterations and apply a *1 – hypergeometric* test to determine whether co-alterations occur significantly less often than expected under independence. Significant underrepresentation is taken as evidence of a synthetic-lethal relationship.
2. **Alteration patterns considered**
The method includes both loss–gain (e.g., BRCA2 loss with TLK2 amplification) and loss–loss (parallel DDR deficiencies) combinations. Treating these together is reasonable because SL may arise either from disrupting complementary pathways or from breaking regulatory feedback loops.
3. **Validation with essentiality screens**
The authors use genome-wide essentiality data from cell lines deficient in gene A. Candidate B genes are consistently more essential in A-defective lines than in proficient controls. This context-specific essentiality provides strong functional support for the SL interpretation.
4. **Biological plausibility**
Several predicted pairs match known interactions (e.g., PIK3CA–PTEN), and many candidate B genes correlate with poorer patient survival when overexpressed, consistent with tumour fitness roles.

Is their test a good one?

Strengths

- Directly leverages human tumour data.
- Statistical test appropriately detects underrepresented double events.
- Essentiality validation strengthens confidence in SL predictions.
- Context comparisons (e.g., PTEN^+/+ vs PTEN^–/–) support causality.

Limitations

- Mutual exclusivity can arise from non-SL causes (subtypes, pathway saturation).
- Limited power for rare alterations.
- Cell-line essentiality may not fully reflect tumour biology.

Overall assessment

Mutual exclusivity alone cannot prove SL, but paired with solid essentiality validation and plausible mechanisms, the approach is well justified and useful for prioritizing SL candidates in tumours.

Homework #3, due 24/1/2026

submit by email to dcswls@nus.edu.sg

Read this paper, focusing on Section 3.1, Segmentation by eigenbackground subtraction:

Oliver et al., “A Bayesian computer vision system for modeling human interactions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831-843, 2000

Read the review report (attached below) produced by Gemini on this section.

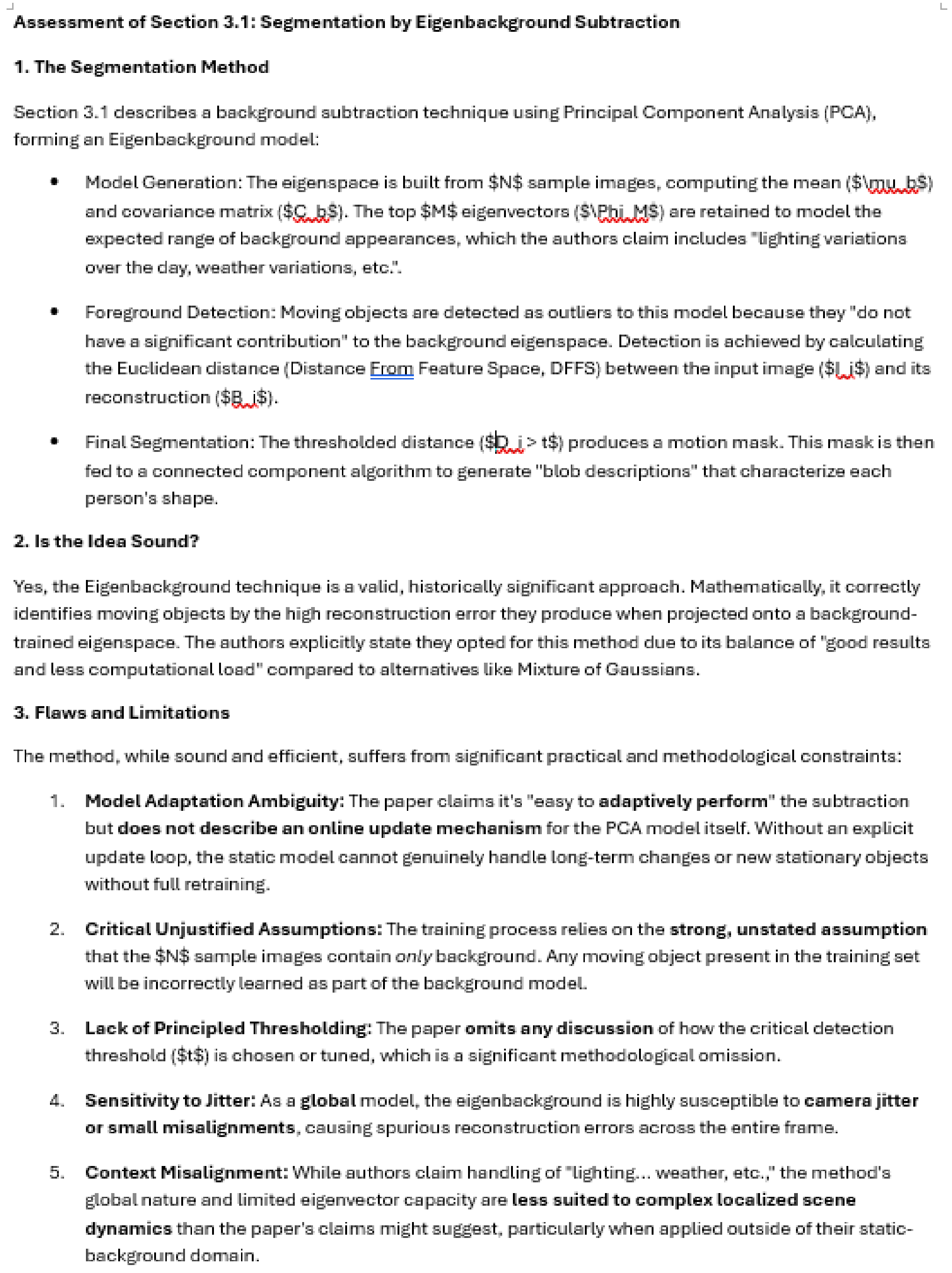
Answer the following questions:

1/ Which parts of the Gemini report, if any, do you agree with?

2/ Which parts of the Gemini report, if any, do you disagree with? Why?

3/ Are there any major flaws that you think should be pointed out but are missed by Gemini?

Make 5-minutes presentation to class on 26/1/2026



Assessment of Section 3.1: Segmentation by Eigenbackground Subtraction

1. The Segmentation Method

Section 3.1 describes a background subtraction technique using Principal Component Analysis (PCA), forming an Eigenbackground model:

- Model Generation: The eigenspace is built from N sample images, computing the mean (μ_b) and covariance matrix (C_b). The top M eigenvectors (Φ_M) are retained to model the expected range of background appearances, which the authors claim includes "lighting variations over the day, weather variations, etc.".
- Foreground Detection: Moving objects are detected as outliers to this model because they "do not have a significant contribution" to the background eigenspace. Detection is achieved by calculating the Euclidean distance (Distance From Feature Space, DFFS) between the input image (I_i) and its reconstruction (B_i).
- Final Segmentation: The thresholded distance ($D_i > t$) produces a motion mask. This mask is then fed to a connected component algorithm to generate "blob descriptions" that characterize each person's shape.

2. Is the Idea Sound?

Yes, the Eigenbackground technique is a valid, historically significant approach. Mathematically, it correctly identifies moving objects by the high reconstruction error they produce when projected onto a background-trained eigenspace. The authors explicitly state they opted for this method due to its balance of "good results and less computational load" compared to alternatives like Mixture of Gaussians.

3. Flaws and Limitations

The method, while sound and efficient, suffers from significant practical and methodological constraints:

1. **Model Adaptation Ambiguity:** The paper claims it's "easy to **adaptively perform**" the subtraction but **does not describe an online update mechanism** for the PCA model itself. Without an explicit update loop, the static model cannot genuinely handle long-term changes or new stationary objects without full retraining.
2. **Critical Unjustified Assumptions:** The training process relies on the **strong, unstated assumption** that the N sample images contain *only* background. Any moving object present in the training set will be incorrectly learned as part of the background model.
3. **Lack of Principled Thresholding:** The paper **omits any discussion** of how the critical detection threshold (t) is chosen or tuned, which is a significant methodological omission.
4. **Sensitivity to Jitter:** As a **global** model, the eigenbackground is highly susceptible to **camera jitter or small misalignments**, causing spurious reconstruction errors across the entire frame.
5. **Context Misalignment:** While authors claim handling of "lighting... weather, etc.," the method's global nature and limited eigenvector capacity are **less suited to complex localized scene dynamics** than the paper's claims might suggest, particularly when applied outside of their static-background domain.