UNIVERSITI MALAYA

ALTERNATIVE ASSESSMENT FOR THE DEGREE OF COMPUTER SCIENCE (APPLIED COMPUTING)

ACADEMIC SESSION 2025/2026 : SEMESTER I

WOA7001 : Advanced Algorithm

Januari 2026 *Time: 1 Week*

INSTRUCTIONS TO CANDIDATES:

1. This is a 1-week individual take-home examination. Your submission must be uploaded by the stated deadline. Late submissions will not be accepted unless supported by documented emergencies.

2. Choose ONE (1) case study only from the two provided:
   - Case Study A – Randomised Algorithms for Sensor Network Anomaly Detection
   - Case Study B – Randomised Algorithms for Genome Sequence Analysis

3. Both case studies have the same structure, difficulty, and mark distribution.

4. Your work must be entirely your own reasoning.
   - You may refer to lecture notes, academic papers, and online resources.
   - Directly copying or relying on AI-generated answers is not permitted.
   - You must demonstrate original thinking, custom reasoning, and scenario-based justification.

5. Page guideline: Maximum 10 pages (excluding references and appendices).

(This question paper consists of 2 case studies on 6 printed pages)

Case Study A – Randomised Algorithms for Sensor Network Anomaly Detection

Introduction

The country of *Terran* operates a nationwide environmental hazard sensor network with 1.2 million IoT sensors deployed across forests, rivers, industrial zones, and urban districts. Each sensor sends one reading per minute, including environmental readings (temperature, gas concentration, particulate matter, etc.), signal confidence and sensor metadata (ID, region).

However, the system faces several issues, including, communication noise, base-calling errors from low-cost sensors, network congestion causing missing readings, as well as adversarial spoofing attempts that inject fake anomaly spikes

Because of these challenges, the National Hazard Intelligence Centre (NHIC) must use randomised algorithms to identify anomalous clusters (sensor groups showing abnormal values), estimate high-frequency anomaly indicators in real time, operate on streaming data, where exact deterministic methods exceed time limits and defend against spoofing and sensor drift.

The goal is to design a randomised, streaming-capable anomaly detection pipeline.

Dataset Fragment and Constraint

Below is a fragment from a single minute of sensor data. The data is incomplete, noisy, and ambiguous.

Sensor	Region	Reading	Signal Confidence	Notes
S201	R2	57	0.91	Normal
S202	R2	141	0.74	High but stable
S203	R2	149	0.18	Likely spoofed
S204	R3	? (missing)	0.88	Connectivity drop
S205	R3	152	0.55	Noise suspected
S206	R3	150	0.93	High reading
S207	R4	31	0.97	Normal
S208	R4	29	0.94	Normal
S209	R2	? (corrupted)	0.12	Very low-quality signal
S210	R3	153	0.61	Sudden spike

Detection Constraints

- Readings >145 are potential anomalies but must be weighted by signal confidence.
- Missing values (?) must be probabilistically inferred, not averaged.
- R2 and R3 are known high-risk zones in which anomalies are more significant.

- Spoofing attempts (e.g., S203) must be considered in your design.
- Streaming throughput requires sublinear-memory algorithms.

Answer the following questions.

1. Propose a randomised detection pipeline for identifying anomalies at scale.

Your answer must use at least *one* of these techniques:
- Random sampling (reservoir or weighted)
- Sketching (Count-Min, AMS, HyperLogLog variants)
- Locality-Sensitive Hashing (LSH)
- Monte Carlo anomaly estimation
- Randomised filtering structures

2. Provide a probabilistic reconstruction of the missing values in S204 and S209 based on:
   - Spatial similarity (region-based clustering)
   - Confidence weights
   - Noise models

3. Compare your randomised approach against two deterministic methods, evaluating:
   - Scalability
   - Error tolerance
   - Real-time suitability
   - Resistance to adversarial inputs

4. Critically evaluate your design under real-world constraints:
   - Accuracy vs efficiency trade-offs
   - Scaling to 1.2 million sensors
   - Behaviour during extreme noise bursts
   - Limitations of your randomised method

Case Study B – Randomised Algorithms for Genome Sequence Analysis

Introduction

The National Genomics Institute of Bayu (NGIB) sequences DNA from a rapidly mutating agricultural crop. Each genome contains >3.2 billion base pairs, but sequencing machines produce noisy, incomplete reads, ambiguous bases ("N" or "?"), fragmented segments, occasional contamination from other species and time-sensitive streams (hundreds of samples/day).

Deterministic algorithms (full alignments, exact counts) are prohibitively slow. NGIB requires randomised, approximate genomic algorithms to estimate high-frequency k-mers (substrings of length k extracted from a DNA sequence), detect mutation hotspots, identify contaminated segments, and process genome data in a streaming fashion with limited memory.

Dataset Fragment and Constraint

Below are three noisy genomic fragments from the same chromosome region:

Fragment	Genome Sequence
A	ATGCC??TAGGCTATANNCTG
B	ATGACCTAGG?TATACACTG
C	GTGCCA?AGGCTATACNCT?

Constraints

- "?" and "N" represent ambiguous nucleotides from sequencing noise.
- A and B appear related; C may belong to a different species.
- Mutation hotspots must be estimated probabilistically.
- Contamination must be detected using similarity estimation rather than full alignment.
- NGIB requires randomised methods due to the massive data size.

Answer the following questions.

1. Propose a randomised pipeline using at least one of:
   - MinHash or LSH for genomic similarity
   - Randomised pattern matching (e.g., Monte Carlo Karp–Rabin)
   - Randomised sampling of k-mers
   - Probabilistic sketches for approximate counting

2. Probabilistic reconstruction of ambiguous bases in A, B, and C
   - Based on mutation likelihoods
   - Comparison across fragments
   - Error modelling

3. Compare your randomised method against two deterministic algorithms for genomic analysis, evaluating:
   - Time
   - Memory
   - Suitability for extremely large genomes
   - Noise tolerance

4. Evaluate your design under large-scale genomic constraints:
   - Memory and runtime effects of sketching
   - False-positive and false-negative risks
   - Streaming throughput considerations
   - Effects of noise concentration

Evaluation Rubrics

Criteria A — Understanding of Randomised Algorithm(s)

Marks	Descriptor
5	Demonstrates comprehensive understanding of chosen randomised techniques with accurate descriptions and correct assumptions.
4	Shows strong understanding with minor conceptual gaps.
3	Adequate but partially generic understanding; some inaccuracies.
2	Limited understanding; superficial explanation of techniques.
1	Fragmented or incorrect understanding; major conceptual errors.
0	No valid explanation provided.

Criteria B — Quality of Justification & Method Selection

Marks	Descriptor
5	Provides convincing, evidence-based justification for algorithm choice; critically compares with at least two alternatives in context of the case study’s constraints.
4	Strong justification with reasonable comparison; some areas could be more precise.
3	General justification; comparison largely generic or superficial.
2	Weak justification; comparison incomplete or flawed.
1	Minimal reasoning; choices not defended.
0	No justification or irrelevant content.

Criteria C — Dataset-Dependent Reasoning

Marks	Descriptor
5	Excellent integration of the dataset; probabilistic reconstruction of missing/ambiguous values is well-reasoned and clearly tied to algorithm design.

4	Good dataset integration with minor inconsistencies.
3	Adequate dataset use but partially generic or incomplete reconstruction.
2	Minimal dataset linkage; reconstruction not grounded in context.
1	Very weak or incorrect reference to dataset.
0	Dataset not used at all.

Criteria D — Critical Evaluation of Large-Scale Constraints

Marks	Descriptor
5	Insightful, detailed, evidence-based evaluation; clearly addresses all required dimensions with depth.
4	Strong evaluation with minor gaps; mostly comprehensive.
3	Adequate evaluation but lacks depth or misses one dimension.
2	Superficial evaluation with limited critical thought.
1	Minimal evaluation; misses several key issues.
0	No evaluation provided.

Criteria E — Academic Structure, Clarity & Referencing

Marks	Descriptor
5	Clear, well-organised, academically written; diagrams/tables enhance understanding.
4	Generally clear with minor structural weaknesses.
3	Understandable but occasionally unclear or unfocused.
2	Disorganised or difficult to follow.
1	Very poor clarity; confusing narrative.
0	Not submitted or unreadable.

Total marks: 25

END