## **Postgraduate Seminar Series**

Venue: Graduate Seminar Room Date & Time: May 24, 2025 at 2:30 PM

## **Speaker Information**

Samin Rahman Khan (Std No. 0422052003) is a part time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies from Bangladesh University of Engineering and Technology (BUET) in 2022. His research interest lies in the fields of Bioinformatics, Machine Learning, and High-Dimensional Statistical Analysis. He is currently doing his postgraduate thesis under the supervision of Dr. Mohammad Saifur Rahman. He will be speaking about his ongoing research in this talk.



## Quantifying Pathological Progression from Single-Cell Transcriptomics Data

The ability to measure cellular state transitions between healthy and diseased conditions is fundamental to understanding disease mechanisms and progression. The increasing availability of single-cell datasets and large-scale reference atlases enables the comparison of cell states across different conditions. However, existing tools lack the capability to identify cell populations that have shifted statistically significantly from a reference state - a critical aspect for accurate disease characterization. To address this need, we introduce single-cell Pathological Shift Scoring (scPSS), a computational method that quantifies the statistical significance of cellular state deviations from a reference. Importantly, scPSS does not require annotated disease datasets, making it uniquely applicable to rare and emerging diseases.

Current methods have improved disease-specific feature identification, they don't provide a quantitative assessment of state alterations and lack the ability to measure the degree of disease-associated changes. Recent machine learning methods have enabled the identification and scoring of disease-relevant cellular states, but their dependence on labeled training data from both healthy and diseased individuals limits their application to well-characterized conditions.

Our approach, scPSS, uses gene expression profiles from normal cells to establish a reference state distribution, using k-nearest neighbor distances in principal component embedding space. For any query cell, it calculates a "pathological shift score" that measures its deviation from this healthy state distribution. This score enables both the ranking of cells by their degree of state deviation and the identification of disease progression. The key differentiator for scPSS is its semi-supervised reference-based design - quantifying pathological shifts using only healthy reference distributions, without requiring training on condition-labeled data. By employing a simple yet robust statistical framework, scPSS provides an interpretable measure of cellular state deviation while adhering to the principle of Occam's Razor.

In comparative analyses across multiple datasets containing healthy and diseased cells, scPSS matched or exceeded the performance of the state-of-the-art Contrastive VI method (modified to provide pathological shift scores), showing substantial improvements in AUPR of up to 20% in most datasets. We also used the scPSS to identify healthy and diseased individuals from the proportion of diseased cells compared with reference datasets. Using the HCLA dataset, scPSS successfully distinguished healthy from IPF conditions with 86% accuracy using only healthy reference cells and 91% accuracy when both healthy and diseased reference cells were available.