Name: **Shahid Ahmad Wani** Notification No: **584/2025**

Name of the Supervisor: **Prof. S. M. K. Quadri** Notification Date: **18-08-2025**

Topic: **Knowledge Discovery with Machine Learning for Multi-modal Single-cell Data**

Department / Faculty: **Computer Science, Faculty of Sciences**

FINDINGS

This thesis introduces how advances in single-cell technologies particularly **scRNA-seq** have revolutionized our capacity to understand individual cell behaviour and diversity. It points out that while these technologies offer tremendous potential, they also introduce analytical challenges like data sparsity, high dimensionality, and batch variation. It advocates for the application of machine learning and deep learning to manage these complexities, emphasizing the importance of robust preprocessing and integration methods. This work also offers a detailed historical and technical review of single-cell sequencing, outlining how the field has evolved from early RNA profiling techniques to today's sophisticated multimodal protocols. It explains the end-to-end pipeline from cell isolation and gene expression capture to computational analyses like clustering, annotation, and trajectory mapping. The integration of deep learning into single-cell analysis, describing how models such as autoencoders, graph networks, and VAEs can overcome the limitations of traditional tools is also explored in this work.

Furthermore, this study centres on understanding how various normalization strategies influence downstream analytical performance in multimodal single-cell datasets. It explores and compares techniques such as log-alpha, acosh, and log-CPM, using datasets like PBMC and SHARE-seq as test cases. The study assesses their effect on data integration and clustering accuracy, supported by visual tools like UMAP and quantitative metrics like ARI and NMI. The study highlights that preprocessing isn't just a routine step, but a crucial determinant of whether subsequent biological patterns are accurately recovered or obscured by noise.

This thesis introduces **scJVAE**, a novel variational autoencoder-based framework developed to jointly embed multimodal single-cell data while mitigating batch effects. Unlike existing methods that treat modalities separately or rely on complex preprocessing, scJVAE learns an integrated latent space using parallel encoders and decoders for RNA and ATAC data. The model is rigorously tested on four datasets of varying size and complexity, demonstrating superior results in clustering fidelity, modality alignment, and memory efficiency when compared to other popular tools. This work showcases **scJVAE** as a scalable and biologically reliable solution for joint data analysis.

Focusing on cell-type classification, this work evaluates the effectiveness of several machine learning models—ranging from basic algorithms like decision trees and KNN to more sophisticated approaches like SVMs and Transformer-based classifiers. Multiple datasets with distinct cellular compositions are used to assess model performance based on accuracy, F1-score, and clustering visuals. The findings indicate that while simpler models offer ease of use,

they struggle with the complexity of real-world data, whereas models with stronger generalization capacity like SVM, Logistic Regression and transformers yield better annotation outcomes. The major findings and innovations of the thesis suggests how the choice of preprocessing, the development of **scJVAE**, and the evaluation of ML methods collectively contribute to improving multimodal single-cell analysis. This work reiterates the importance of robust integration and accurate cell labelling in uncovering biological insights.