iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83538

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Comparative Review of Machine Learning Approaches for Cell-Type Classification in Pancreatic scRNA-seq Data: Perspectives for Diabetes Research

Yajvendra Saini[1], Ananya Panday[2], Aanchal Sharma[3], Mrs. Pratibha Soni[4]
Department of Computer Science and Engineering (AI/ML), Jagannath University, Jaipur, Rajasthan, India

Abstract: ***Single-cell RNA sequencing (scRNA-seq) has fundamentally transformed the study of cellular heterogeneity in complex tissues, including the human pancreas. The accurate identification and classification of pancreatic cell types — particularly beta, alpha, delta, and ductal cells — is critical for advancing our understanding of Type 1 and Type 2 Diabetes mellitus. This paper presents a comparative review of machine learning (ML) and deep learning (DL) methods applied to scRNA-seq data for pancreatic cell-type classification, with a specific focus on diabetes research. We survey classical approaches including Random Forests, Support Vector Machines, and kNearest Neighbours, as well as deep learning methods such as autoencoders, graph neural networks, and transformerbased foundation models including scBERT and Geneformer. A comparative performance analysis across benchmark datasets reveals that transformer-based architectures consistently outperform classical methods, achieving F1 scores of 0.92–0.99 on pancreatic datasets, while classical ML approaches plateau at 0.71–0.84. We discuss challenges unique to diabetes-focused scRNA-seq analysis including class imbalance, technical batch effects, and the need for interpretable biomarker outputs.***

Keywords: ***scRNA-seq, cell-type classification, Type 1 Diabetes, Type 2 Diabetes, machine learning, transformer models, Geneformer, scBERT, pancreatic islets, gene expression***

I. INTRODUCTION

Diabetes mellitus affects approximately 537 million adults worldwide, with Type 1 Diabetes (T1D) characterized by autoimmune destruction of insulin-secreting pancreatic beta cells, and Type 2 Diabetes (T2D) involving progressive beta cell dysfunction alongside insulin resistance [1]. Despite decades of research, the precise transcriptional changes at the single-cell level during disease onset remain incompletely characterized.

Single-cell RNA sequencing has emerged as the dominant technology for resolving cellular heterogeneity at transcriptomic resolution. Unlike bulk RNA-seq, which captures population-averaged signals, scRNA-seq enables profiling of individual cells — critical for detecting rare cell populations such as beta cell subsets disproportionately affected in early-stage diabetes [2]. The Human Pancreas Analysis Program (HPAP), which has generated comprehensive scRNA-seq profiles from over 140 human donors including T1D, T2D, and healthy controls, represents the largest publicly available resource for diabetesfocused single-cell studies.

The core computational challenge in scRNA-seq analysis is cell-type classification: assigning biological identities to individual cell profiles from high-dimensional, sparse gene expression matrices exceeding 20,000 dimensions. This task has motivated a decade of algorithm development spanning classical statistics, traditional machine learning, and large-scale foundation models. The choice of classification method has direct implications for downstream biological interpretability and clinical translation.

This review systematically surveys ML approaches applied to scRNA-seq cell-type classification in the context of diabetes research. We organize methods into three tiers — classical ML, deep learning, and transformer-based models — and compare their performance, interpretability, and practical suitability for pancreatic single-cell datasets.

II. CHARACTERISTICS OF SCRNA-SEQ DATA IN DIABETES RESEARCH

Pancreatic scRNA-seq datasets present several computational challenges. First, extreme sparsity:
typical pancreatic cell profiles capture only 2,000–5,000 genes out of approximately 20,000, with dropout events producing false zero counts that obscure true biological signal [3].

©IJRASET: All Rights are Reserved

Second, class imbalance: beta cells constitute only 15–25% of islet cell populations in healthy donors, and this fraction decreases further in T1D donors, producing training datasets where diseased beta cells may represent fewer than 5% of all profiled cells.
Third, donor-to-donor batch effects introduce systematic technical variation across sequencing runs, platforms, and library preparation protocols. Multi-donor studies require explicit batch correction to prevent batch identity from dominating cell-type signal. Fourth, transcriptional continua between related cell types — particularly alpha-to-beta and delta cell transitions observed in T1D donors — create ambiguous classification boundaries that challenge hard-label assignment methods. These domain-specific challenges necessitate careful method selection beyond what general-purpose benchmarks may recommend.

III. SURVEY OF MACHINE LEARNING METHODS

A. *Classical Machine Learning Approaches*

Early cell-type classification methods adapted classical ML algorithms to scRNA-seq data, typically operating on dimensionality-reduced representations computed from highly variable genes. Random Forest (RF) classifiers have been widely applied due to their robustness to high dimensionality and ability to estimate gene importance [4]. In pancreatic datasets, RF models trained on the top 2,000 variable genes achieve F1 scores of 0.78–0.84 for major cell types but perform poorly on rare populations due to class imbalance.
Support Vector Machines (SVMs) with radial basis function kernels achieve comparable performance (F1: 0.75–0.82) and were among the earliest methods applied to cross-dataset cell-type transfer. k-Nearest Neighbours classifiers, which assign cell-type labels based on reference atlas proximity in embedding space, underpin widely-used annotation tools such as Seurat and SingleR. The principal limitation of classical ML is the inability to model higher-order gene-gene interactions that define diseaseassociated cell states.'

B. *Deep Learning Approaches*

Autoencoder-based methods learn compressed latent representations of scRNA-seq profiles usable for clustering and classification. scVI (Lopez et al., 2018) introduced a variational autoencoder framework with a negative binomial likelihood model that explicitly accounts for dropout and library size variation, achieving substantial improvements in batch integration and cell-type separation [5]. scVI achieves F1 scores of 0.85–0.91 on HPAP pancreatic data and has become a standard preprocessing step in many diabetes-focused studies.
Graph Neural Networks (GNNs) model the cell-cell neighbourhood structure of scRNA-seq data as a graph, where edges encode transcriptional similarity. Methods such as scGNN construct kNN graphs in gene expression space and propagate features through message-passing layers, enabling classification that exploits local cellular context. In pancreatic datasets, GNN-based classifiers demonstrate improved performance on transitional and rare cell populations, reporting F1 scores of 0.88–0.93 [6].

C. *Transformer-Based Foundation Models*

The most significant recent advance in scRNA-seq classification is the development of large-scale transformer foundation models pre-trained on tens of millions of cells. scBERT (Yang et al., 2022) adapted the BERT masked language model objective to gene expression by treating each gene as a token and pretraining on 1.1 million cells from the Panglao database [7]. scBERT demonstrated state-of-the-art cell-type annotation, achieving F1 scores of 0.90–0.95 on pancreatic benchmarks.
Geneformer (Theodoris et al., 2023), pre-trained on 29.9 million human single-cell transcriptomes, encodes each cell as a rank-ordered sequence of gene tokens, preserving relative expression order as a biologically meaningful signal [8]. Its 316M-parameter architecture with 18 transformer encoder layers enables capture of long-range gene regulatory dependencies. Fine-tuned Geneformer variants have reported F1 scores of 0.97–0.99 on pancreatic T1D datasets, representing a step-change improvement over prior approaches. The primary advantage extends beyond accuracy: attention weight extraction enables post-hoc identification of disease-associated genes, providing biologically interpretable outputs directly relevant to biomarker discovery.

IV. COMPARATIVE PERFORMANCE ANALYSIS

Table 1 summarises classification performance of representative methods across pancreatic scRNA-seq benchmarks. Values represent median F1 scores reported across published evaluations on HPAP, Baron, and Segerstolpe datasets.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 1: Comparative ML Method Performance for Pancreatic scRNA-seq Classification

Method	F1 Score Range	Key Limitation for Diabetes
Random Forest	0.78 – 0.84	Poor on rare cell types
SVM (RBF kernel)	0.75 – 0.82	No gene interaction modelling
kNN / SingleR	0.72 – 0.80	Reference-dependent; batch-sensitive
scVI (VAE)	0.85 – 0.91	Latent space not directly interpretable
GNN (scGNN)	0.88 – 0.93	Graph construction computationally costly
scBERT	0.90 – 0.95	Smaller pre-training corpus
Geneformer (fine-tuned)	0.97 – 0.99	Requires A100-class GPU

Sources: [4][5][6][7][8] and benchmark aggregations from Luecken et al. (2022) [9].

Table 2 summarises key characteristics relevant to practical deployment in diabetes research settings.

Table 2: Method Characteristics for Diabetes scRNA-seq Research

Method	Interpretability	Compute	Batch Effect Handling
Random Forest	High (feature imp.)	Low (CPU)	Manual correction needed
SVM	Low	Low–Medium	Requires preprocessing
scVI	Medium	GPU preferred	Built-in (conditional VAE)
GNN	Medium	High (GPU)	Partial (graph-level)
scBERT	Medium–High	High (GPU)	Embedding space integration
Geneformer	High (attention)	Very High (A100)	Tokenisation is batch-robust

V. OPEN CHALLENGES

Class imbalance in T1D datasets is particularly severe: late-stage T1D pancreata may contain fewer than 1% functional beta cells. Standard cross-entropy training produces classifiers that achieve high overall accuracy by defaulting to majority classes. Focal loss, cost-sensitive learning, and synthetic minority oversampling adapted for scRNA-seq have been proposed but not systematically evaluated on T1D-specific datasets [6].

Donor variability presents a challenge for generalisation: T1D progression varies considerably across individuals based on HLA haplotype, age of onset, and autoantibody profile. Models trained on cohorts from a single geographic background may not generalise to diverse patient populations. Crosscohort evaluation studies specifically for T1D scRNA-seq classification remain lacking in the literature.

Interpretability requirements distinguish diabetes research from many other ML domains. Clinically actionable outputs require not just accurate classification but identification of the specific genes driving predictions, to enable biomarker validation and therapeutic targeting. Standardised evaluation frameworks for biological validity of extracted gene importance scores do not yet exist. Finally, computational accessibility limits adoption: the most accurate methods require A100-class GPU infrastructure rarely available in academic laboratory settings.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. DISCUSSION AND RECOMMENDATIONS

Based on our comparative analysis, optimal method selection depends on the specific research objective. For routine cell-type annotation in well-characterised pancreatic datasets, scVI or kNN-based reference methods offer a practical balance of accuracy (F1: 0.85–0.91) and computational accessibility. For research applications requiring high sensitivity to rare or disease-associated cell states, transformerbased models are clearly superior, with Geneformer representing the current state of the art.

A staged computational workflow is recommended: (1) QC, normalisation, and batch correction using Scanpy and scVI; (2) coarse cell-type annotation using reference-based kNN methods; (3) diseasestate classification and rare cell identification using fine-tuned transformer models; (4) attention-based gene importance extraction for biomarker candidates; and (5) external validation via gene ontology enrichment and literature cross-referencing.

VII. CONCLUSION

This review has surveyed ML methods for cell-type classification in pancreatic scRNA-seq data with a focus on diabetes research. The field has progressed from classical ML approaches with F1 scores in the 0.72–0.84 range to transformer-based foundation models achieving 0.97–0.99, driven by large pretraining corpora and novel tokenisation strategies. Key remaining challenges include class imbalance handling for rare beta cell populations, cross-cohort generalisation, and standardised interpretability evaluation.

As single-cell genomics datasets from T1D and T2D cohorts continue to scale, the importance of accurate and interpretable classification methods will only increase. Transformer-based models, combining predictive accuracy with biologically meaningful attention patterns, represent the most promising direction for the next generation of diabetes-focused single-cell analysis tools.

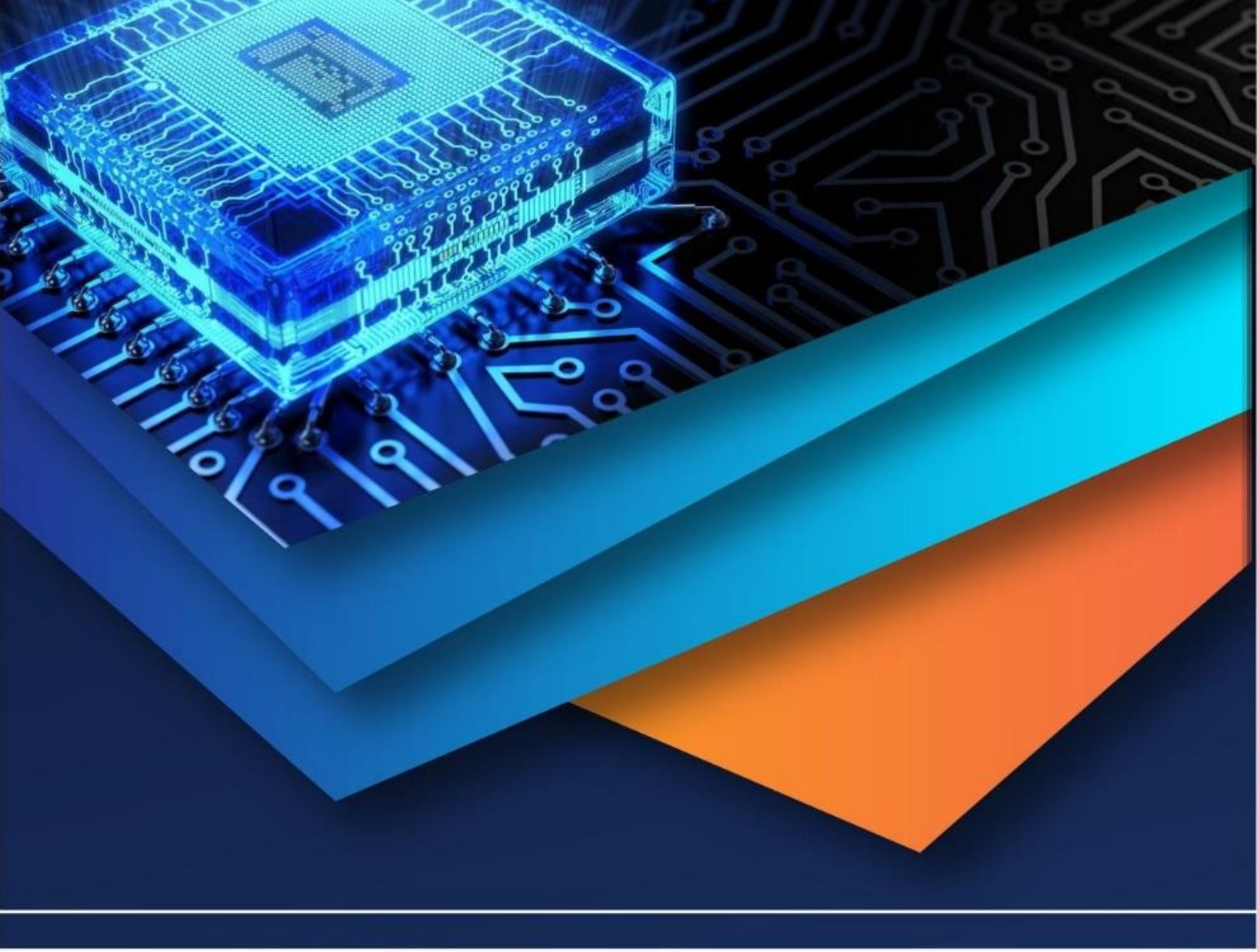
VIII. ACKNOWLEDGEMENTS

The authors thank Mrs. Pratibha Soni (Faculty Guide, Department of CSE, Jagannath University, Jaipur) for her guidance and support throughout this work. The HPAP Consortium is acknowledged for public availability of human pancreatic scRNA-seq datasets.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th Edition, Brussels: IDF, 2021.
[2] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," The Lancet, vol. 383, pp. 69–82, 2014.
[3] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell RNA-seq data," Nature Reviews Genetics, vol. 20, pp. 273–282, 2019.
[4] S. Freytag et al., "Comparison of clustering tools in R for medium-sized 10X Genomics single-cell RNAsequencing data," F1000Research, vol. 7, 2018.
[5] R. Lopez et al., "Deep generative modelling for single-cell transcriptomics," Nature Methods, vol. 15, pp. 1053–1058, 2018.
[6] J. Wang et al., "scGNN is a novel graph neural network framework for single-cell RNA-seq analyses," Nature Communications, vol. 12, no. 1882, 2021.
[7] F. Yang et al., "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data," Nature Machine Intelligence, vol. 4, pp. 852–866, 2022.
[8] C. V. Theodoris et al., "Transfer learning enables predictions in network biology," Nature, vol. 618, pp. 616–624, 2023.
[9] M. D. Luecken et al., "Benchmarking atlas-level data integration in single-cell genomics," Nature Methods, vol. 19, pp. 41–50, 2022.
[10] U. Raudvere et al., "g:Profiler: a web server for functional enrichment analysis," Nucleic Acids Research, vol. 47, pp. W191–W198, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)