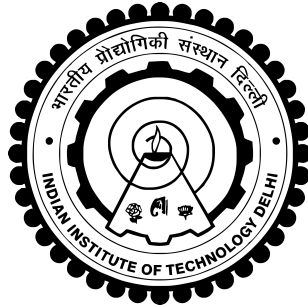


NEIGHBORHOOD DENSITY ESTIMATION USING  
SPACE-PARTITIONING BASED HASHING SCHEMES

AASHI JINDAL



DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY DELHI

June 2023

©Indian Institute of Technology Delhi (IITD), New Delhi, 2023

NEIGHBORHOOD DENSITY ESTIMATION USING  
SPACE-PARTITIONING BASED HASHING SCHEMES

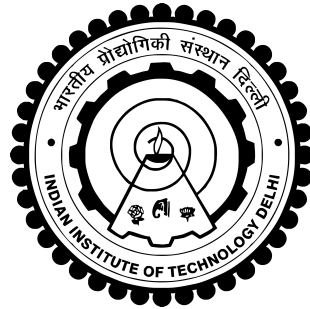
by

AASHI JINDAL

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

*in fulfillment of the requirements of the degree of Doctor of Philosophy  
to the*

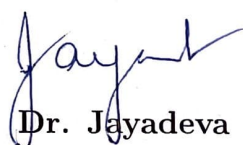


INDIAN INSTITUTE OF TECHNOLOGY DELHI

June 2023

# Certificate

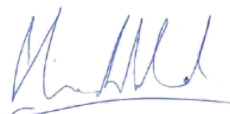
This is to certify that the thesis entitled “Neighborhood Density Estimation Using Space-Partitioning Based Hashing Schemes”, being submitted by Aashi Jindal for the award of the degree of **Doctor of Philosophy** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, is a record of bonafide work done by her under my supervision and guidance. The matter embodied in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma.



**Dr. Jayadeva**

*Professor*

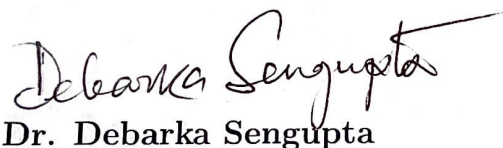
Department of Electrical Engineering,  
Indian Institute of Technology Delhi,  
Hauz Khas, New Delhi - 110016,  
INDIA.



**Dr. Shyam Prabhakar**

*Associate Director*

Laboratory of Systems Biology and Data Analytics,  
Genome Institute of Singapore,  
SINGAPORE



**Dr. Debarka Sengupta**

*Associate Professor*

Department of Computer Science & Engineering,  
Department of Computational Biology,  
Centre for Artificial Intelligence,  
Indraprastha Institute of Information Technology,  
Okhla Phase III, Delhi - 110020, India  
*(Adj.) Associate Professor*  
Institute of Health & Biomedical Innovation,  
QUT, AUSTRALIA

# Acknowledgements

I sincerely thank my supervisors Dr. Jayadeva, Dr. Debarka Sengupta and Dr. Shyam Prabhakar for their valuable feedback. I also thank other SRC members Dr. I.N. Kar, Dr. Shouri Chatterjee and Dr. Vivekanandan Perumal. I thank Prof. Suresh Chandra for his encouragement. I thank my colleagues Dr. Sumit Soman and Dr. Udit Kumar for their moral support. I thank department staff members Rakesh, Yatindra, Satish and Mukesh for always helping with the department work.

I thank my parents, Rakesh Jindal and Meena Jindal, for always believing in me and supporting me in my decisions. I thank my parents-in-law, Sanjiv Kumar Gupta and Snehlata Gupta for being always understanding. I thank my husband Prashant Gupta for always being there whenever I have needed him. He supported me both professionally and personally. My brother, Sahil Jindal always holds a special place in my life. My core family members, Shruti Singla, Praveen Singla, Shaffi Bindal, Mohit Bindal, Khushboo Mehrotra, Yuvan Singla, Tejas Bindal and Vanya Bindal are an important part of my life.

I also thank my current employers, Ashok Juneja and Shweta Singla for giving me the time to complete my pending research work when needed. I am thankful to my company, Applied Solar Technologies India Pvt. Ltd. and my colleagues, Subhdip Rakshit, Amod Kumar, Shubham Sinha, Pulkit Tyagi, Bikas Gupta, Sukrit Singh Negi, Abhishek Chaudhary and Anil Hansda.

The following have been the three main philosophies of my life, "There is an almighty somewhere looking after us and guiding us to the correct path." "Behind every successful woman is her parents, husband and family that supported her in her unconventional choices." Last, but not the least, "...all's well, that ends well."



*(Aashi Jindal)*

## Abstract

Single cell messenger RNA sequencing (scRNA-seq) offers a view into transcriptional landscapes in complex tissues. Recent developments in droplet based transcriptomics platforms have made it possible to simultaneously screen hundreds of thousands of cells. It is advantageous to use large-scale single cell transcriptomics since it could lead to the discovery of a number of rare cell sub-populations. When the sample size reaches the order of hundreds of thousands, existing techniques to discover rare cells either scale unbearably slow or terminate altogether. We suggest the Finder of Rare Entities (FiRE), an algorithm that quickly assigns a rareness score to every individual expression profile under consideration. We show how FiRE scores can assist bioinformaticians in limiting the downstream analyses to only on a subset of expression profiles within ultra-large scRNA-seq data.

Anomaly detection methods differ in their time complexity, sensitivity to data dimensions, and their ability to detect local/global outliers. The proposed algorithm FiRE is a 'sketching' based linear-time algorithm for identifying global outliers. FiRE.1, an extended implementation of FiRE fares well on local outliers as well. We provide an extensive comparison with 18 state-of-the-art anomaly detection algorithms on a diverse collection of 1000 annotated datasets. Five different evaluation metrics have been employed. FiRE.1's performance was particularly remarkable on datasets featuring a large number of local outliers. In the sequel, we propose a new "outlierness" criterion to infer the local or global identity of outliers.

We propose Enhash, a fast ensemble learner that detects *concept drift* in a data stream. A stream may consist of abrupt, gradual, virtual, or recurring events, or a mixture of various types of drift. Enhash employs projection hash to insert an incoming sample. We show empirically that the proposed method has competitive performance to existing ensemble learners in much lesser time. Also, Enhash has moderate resource requirements. Experiments relevant to performance comparison were performed on 6 artificial and 4 real datasets consisting of various types of drifts.

एकक कोशिका दूत-RNA (scRNA-seq) अनुक्रमण जटिल ऊतकों के लिप्यंतरण का एक दृश्य प्रस्तुत करता है। छोटी बूंदों पर आधारित अत्याधुनिक प्रतिलेख यंत्रों के माध्यम से लाखों एकक कोशिकाओं को एक साथ परखना संभव हो गया है। बड़े पैमाने पर एकक कोशिकाओं के प्रतिलेखों का उपयोग दुर्लभ कोशिकाओं को खोजने में किया जा सकता है। जब एकक कोशिकाओं के प्रतिदर्शों का आकड़ा लाखों में पहुँचता है तो उपलब्ध कलन विधिओं असहनीय रूप से शिथिल हो जाती हैं तथा इनका उपयोग करके दुर्लभ कोशिकाओं की खोज करना लगभग असंभव हो जाता है। इस शोध प्रबंध में दुर्लभता के अंकमान की गड़ना के लिए एक कलन विधि, दुर्लभ इकाईओं का खोजक (FiRE) को प्रस्तुत किया गया है। यह कलन विधि तीव्र गति से प्रत्येक कोशिकाओं के प्रतिलेख को एक दुर्लभता का अंकमान प्रदान करती है। हम यह भी दिखाते हैं FiRE अंकमान जैव सूचना वैज्ञानिकों को अत्यधिक विशाल scRNA-seq आंकड़ाकोशों में से केवल कुछ महत्वपूर्ण एकक कोशिकाओं के विश्लेषण करने में सहायता करता है।

विसंगति का पता लगाने के तरीके उनकी समय जटिलता, विवरण आयामों, तथा स्थानीय/वैश्विक रूप से विचित्र प्रतिदृश्यों की खोज करने की छमता में भिन्न होते हैं। प्रस्तावित कलन विधि FiRE वैश्विक रूप से विचित्र प्रतिदृश्यों को रेखीय अवधि में अवगत कराती है तथा इसके लिए स्केचिंग, एक हैशिंग पे आधारिक कलन विधि का उपयोग करती है। FiRE का विस्तारित कार्यान्वयन FiRE.1 स्थानीय रूप से विचित्र प्रतिदृश्यों की खोज करने में सक्षम है। इस शोध प्रबंध में हम १००० मापदंड आंकड़ा संचयों के विविध संग्रह पे १८ अत्याधुनिक विसंगति का पता लगाने वाली कलन विधिओं की व्यापक तुलना प्रदान करते हैं। इस तुलना के लिए ५ भिन्न प्रकर की तुलना विधिओं को नियोजित किया गया है। बड़ी संख्या में स्थानीय रूप से विचित्र प्रतिदृश्यों से परिपूर्ण आंकड़ा संकलनो पर FiRE.1 का प्रदर्शन उल्लेखनीय है। अगली कड़ी में हम स्थानीय और वैश्विक रूप से विचित्र प्रतिदृश्यों का अनुमान लगाने के लिए एक "बाह्यता" मानदंड प्रस्तावित करते हैं।

इस शोध प्रबंध के अंतिम भाग में हम एक तीव्र समवेत कलन विधि, Enhash, को प्रस्तुत करते हैं जो आंकड़ों की धारा में से अवधारणा के बहाव का पता लगाने में सक्षम है। एक आंकड़ों की धारा में आकस्मिक, क्रमिक, आभासी, या आवर्ती घटनाये, या इनके मिश्रण उपास्थिक हो सकते हैं। Enhash आने वाले प्रतिदर्शियों को सम्मलित करने के लिए प्रक्षेप हैश का उपयोग करता है। हम अनुभवजन्य रूप से दिखाते हैं की प्रस्तावित कलन विधि का प्रदर्शन उपलब्ध समवेत आधारित कलन विधिओं की अपेक्षा प्रतिस्पर्धी है तथा प्रस्तावित कलन विधि यह प्रदर्शन अन्य विधिओं की तुलना में कम समय तथा मध्यम संसाधनों का उपयोग करके प्राप्त करती है। प्रस्तावित कलन विधि के प्रदर्शन की तुलना ६ कृत्रिम तथा ४ वास्तविक आंकड़ा संचयों पर की गई है। यह आंकड़ा संचय विभिन्न प्रकार के अवधारणाओं के मिश्रण से बने हैं।

# Contents

<b>Certificate</b>	<b>i</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Hindi Abstract</b>	<b>vii</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xix</b>
<b>List of Abbreviations</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Scope and objectives . . . . .	1
1.2 Locality Sensitive Hashing . . . . .	2
1.2.1 Sketching . . . . .	3
1.2.2 Binary hash . . . . .	3
1.2.3 Projection hash . . . . .	3
1.3 Outlier detection . . . . .	4
1.3.1 Outlier detection: Definition . . . . .	4
1.3.2 Outlier detection: Use cases . . . . .	4
1.3.3 Outlier detection: Existing techniques . . . . .	4
1.3.4 Outlier detection: Motivation . . . . .	9
1.3.5 Outlier detection: Proposed technique . . . . .	10

1.4	Concept drift detection . . . . .	11
1.4.1	Concept drift detection: Definition and use cases . . . . .	11
1.4.2	Concept drift detection: Existing techniques . . . . .	11
1.4.3	Concept drift detection: Motivation and proposed technique . . . . .	12
1.5	Organization of the thesis . . . . .	13
<b>2</b>	<b>Identification of Rare Events*</b>	<b>15</b>
2.1	Introduction . . . . .	15
2.2	Overview of FiRE . . . . .	17
2.2.1	Steps involved in FiRE . . . . .	19
2.3	Results . . . . .	20
2.3.1	Experimental setup . . . . .	20
2.3.2	FiRE discovers cells with varying degrees of rareness . . . . .	25
2.3.3	FiRE detects artificially planted rare cells with high accuracy . . . . .	26
2.3.4	FiRE is sensitive to cell type identity . . . . .	27
2.3.5	FiRE is scalable and fast . . . . .	29
2.3.6	FiRE resolves heterogeneity among dendritic cells . . . . .	33
2.4	Conclusions . . . . .	35
<b>3</b>	<b>Linear Time Identification of Local and Global Outliers<sup>†</sup></b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	FiRE.1: FiRE for local outliers . . . . .	38
3.3	Outlier score on unseen data . . . . .	41
3.3.1	Local vs Global outliers . . . . .	42
3.4	Run time complexity . . . . .	42
3.5	Experimental setup . . . . .	44
3.5.1	Description of datasets . . . . .	44
3.5.2	Metrics for comparing outlier detection methods . . . . .	44
3.6	Performance comparison of methods on a repository of almost 1000 datasets . . . . .	46
3.6.1	Tuning of hyperparameters . . . . .	46

---

\*The work presented in this chapter has been published as a research paper titled “*Discovery of rare cells from voluminous single cell expression data*” in Nature Communications (2018).

<sup>†</sup>The work presented in this chapter has been published as a research paper titled “*Linear time identification of local and global outliers*” in Neurocomputing (2021).

3.6.2	Comparison of methods using Friedman ranking . . . . .	47
3.6.3	Comparison of linear complexity methods from the perspective of outlierness type . . . . .	49
3.6.4	Performance comparison of linear-time methods on large datasets	51
3.7	Conclusions . . . . .	53
<b>4</b>	<b>Enhash: A Fast Streaming Algorithm for Concept Drift Detection<sup>‡</sup></b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	The proposed method: Enhash . . . . .	56
4.2.1	Implementation details . . . . .	59
4.2.2	Time complexity analysis . . . . .	61
4.3	Experimental Setup . . . . .	64
4.3.1	Evaluation metrics for performance comparison . . . . .	64
4.3.2	Description of datasets . . . . .	64
4.3.3	System details . . . . .	65
4.4	Tuning of parameters for Enhash . . . . .	65
4.4.1	Constraints to tune L . . . . .	66
4.4.2	Constraints to tune <i>bin-width</i> . . . . .	66
4.5	Experimental Results . . . . .	67
4.6	Ablation study . . . . .	71
4.7	Conclusions . . . . .	72
<b>5</b>	<b>Conclusions and Future Work</b>	<b>73</b>
5.1	Conclusions . . . . .	73
5.2	Future Work . . . . .	75
5.2.1	Identification of anomalies in time-series data . . . . .	75
	<b>List of Publications</b>	<b>91</b>
	<b>Brief Biodata of Author</b>	<b>93</b>

---

<sup>‡</sup>The work presented in this chapter has been published as a research paper titled “*Enhash: A fast streaming algorithm for concept drift detection*” in ESANN proceedings (2021).

# List of Figures

1.1	Overall categorization of well-known unsupervised anomaly detection algorithms. The three broad categories are statistical, sub-space based, and nearest-neighbor based. . . . .	9
2.1	Overview of FiRE. The first step is to assign each cell to a hash-code. As numerous similar cells can share the same hash-code, it is possible to think of a hash-code as an imagined bucket. The phase of creating the hash-code is repeated $L$ times to test the reliability of rarity estimates. The chance that any point will fall into the bucket of a given cell $i$ and estimator $l$ is calculated as $p_{il}$ . These probabilities are combined in the algorithm's second phase to get an estimate of how rare each cell is. . .	18
2.2	Stability of FiRE. (a),(c) RMS (Root Mean Square) difference in values of FiRE-score of every cell between two successive estimators. For calculation of RMS, FiRE-score is averaged across multiple seeds and normalized by the value of $L$ . (b),(d) RMS difference in values of FiRE-score between two successive values of $M$ . For calculation of RMS, FiRE-score is averaged across multiple seeds and normalized by the value of $M$ . (a)-(b) RMS has been shown on a simulated dataset consisting of a mixture of Jurkat and 293T cells [148]. (c)-(d) RMS has been shown on $\sim 68k$ Peripheral Blood Mononuclear Cells (PBMCs) [148]. . . . .	24

2.3	Performance evaluation of FiRE on Peripheral Blood Mononuclear Cells (PBMCs). (a) t-SNE based 2D embedding of the data with color coded cluster identities as reported by Zheng and colleagues [148]. (b) Rare population identified by FiRE using IQR-thresholding-criteria. (c) Heat map of FiRE scores for the individual PBMCs. The cluster of megakaryocytes (0.3%), the rarest of all the cell types are assigned the highest FiRE scores.	25
2.4	In the $\sim 68k$ PBMC data [148], the appearance of minor cell populations with varying degrees of rarity is accompanied by a rise in the number of chosen rare cells. Figures (a)-(c) demonstrate, respectively, the top 0.25%, 2%, and 5% cells chosen based on FiRE scores. . . . .	26
2.5	Minor cell types' detectability in a simulated dataset with a mixture of Jurkat and 293T cells (known annotations) [148]. (a) $F_1$ scores were determined relative to the rare (Jurkat) population, while bioinformatically altering the percentage of artificially planted rare cells. It is noteworthy that both FiRE and LOF [13] use a threshold to their continuous scores for zeroing on the rare cells. On the other hand, GiniClust [63] and RaceID [51] offer binary annotations for cell-rarity. (b) t-SNE based 2D embedding of the cells with color-coded identities. (c) FiRE-score intensities were displayed on the t-SNE based 2D map. Figures (d)-(g) demonstrate the rare cells detected by various algorithms. (h) Congruence of methods with known annotations. Note: Results shown in Figures (b)-(h) correspond to a rare cell concentration of 2.5%. . . . .	28
2.6	Congruence of methods with known annotations and congruence between pairs of methods on a simulated, scRNA-seq data consisting of 293T and Jurkat cells mixed <i>in vitro</i> in equal proportion [148]. . . . .	29

2.7	<p>Congruence of methods using Venn diagrams. (a),(b) Performance comparison of FiRE, GiniClust [63], RaceID [51] and LOF [13] as per the rare cells identified by them. (a) Performance comparison of methods above on Embryonic Stem Cells (ESCs) data [70]. FiRE could easily identify the Zscan-4 enriched, 2C-like cell cluster as reported by Jiang <i>et al.</i> [63]. Also, the FiRE predicted rare cells had the least overlap with the ones predicted by RaceID, which could not identify those 2C-like cells. (b) Performance on mouse small intestine cells [51]. FiRE could identify the rare cell types in the secretory lineage, which consisted of goblet, enteroendocrine, paneth and tuft cells (as discussed in Grun <i>et al.</i> [51]).</p>	30
2.8	<p>Sensitivity of FiRE to cell type. As soon as there are enough differentially expressed genes to create a small cluster that represents the minor cell subpopulation, fire begins properly identifying the minor cell type on scRNA-seq data generated using the R tool <code>splatter</code> [143]. The succeeding ROC-AUC plots use the figure in the upper-left corner as their legend. Each t-SNE and ROC figure pair represents one of the 1000 times the experiment was run with respect to a particular set of differentially expressed genes. Cell-group annotations were used in the ROC-AUC study, and individual cells were given FiRE ratings. . . . .</p>	31
2.9	<p>FiRE is fast. Execution time collected for the four methods with cell counts varying from 1k to ~68k. . . . .</p>	32
2.10	<p>FiRE-defined dendritic cell heterogeneity in human blood. (a) t-SNE based 2D plot of rare cells detected by FiRE. Cells are color coded based on their cluster identity as determined by dropClust. (b) Dendritic cells, annotated by the authors, are highlighted in the 2D map adopted from Zheng <i>et al</i> [148]. (c) Dendritic cell sub-types detected from FiRE-reported rare cells are color-coded as per Figure (a). (d) Characterization of dendritic cell sub-types using markers reported by Villani and colleagues [124]. . . . .</p>	34
2.11	<p>2D embedding of rare cells detected by FiRE on ~68k PBMCs. Cells are color coded based on the cell type annotations reported by Zheng [148].</p>	35

3.1	Performance of FiRE and FiRE.1 on a simulated dataset with local and global outliers (the illustration is inspired by [79]). (a) A 2-dimensional simulated dataset containing both local and global outliers. (b) Distribution of <i>FiRE-score</i> when $M = 2$ . Global outliers are well captured and have high values of <i>FiRE-score</i> . On the other hand, local outliers deviating marginally from a local population are not captured. (c) Distribution of <i>FiRE-score</i> when $M > d$ ( $M = 100$ ). In addition to global outliers, local ones are also identified since sufficient hash indexes account for minor differences between local outliers and their local population. (d) FiRE.1 identifies both local and global outliers. (e) Variation of outlieriness criterion <i>o-score</i> on different types of outliers. For global outliers, the values are high and vice-versa. . . . .	39
3.2	An overview of FiRE.1 on a simulated dataset. The heatmap depicts FiRE.1 approximated regional densities. . . . .	40
3.3	Performance comparison of 18 outlier detection methods on $\sim 1000$ datasets. The performance of all methods was evaluated on 5 evaluation metrics. Friedman ranking determined for every method for each metric. The method with the smallest Friedman rank performs the best on a given measure. The color intensity in a heatmap depicts the inverse of the Friedman ranking. . . . .	48
3.4	Performance comparison of linear time complexity outlier detection methods - FiRE.1, FiRE, and HBOS based on <i>o-score</i> . The datasets are grouped into 5 different clusters using hierarchical clustering. Cluster#1 consists of 127 (11.2% of the entire collection of datasets), cluster#2 has 667 (59.1%), cluster#3 has 31 (2.7%), cluster#4 has 113 (10%), and cluster#5 has 191 (16.9%) datasets. Every row in the heatmap represents a dataset. The <i>o-score</i> of a given dataset is distributed in 20 bins of a histogram. The bin edges are arranged in columns along the corresponding row of the dataset. The distribution of <i>o-score</i> varies across clusters. . . . .	50

3.5	Density plot and heatmap illustrating the distribution of <i>o-score</i> for 4 different large size datasets. The density plot shows the frequency distribution of <i>o-score</i> for outliers. Every row in the heatmap represents a dataset. The <i>o-score</i> of a given dataset is distributed in 20 bins of a histogram. The bin edges are arranged in columns along the corresponding row of the dataset. . . . .	52
4.1	Enhash accommodates both virtual and real drift. . . . .	58
4.2	Tuning of <i>L</i> . For both synthetic and real datasets, we show a trend in performance metrics, Error (%), and Time(hrs) for an increase in values of <i>L</i> . The value of <i>L</i> varies from [2, 14]. For a given value of <i>L</i> , the running time is measured across all the samples for all the estimators in a configuration. The value of error is evaluated across all the samples for a given value of <i>L</i> . . . . .	66
4.3	Tuning of <i>bin-width</i> . For both synthetic and real datasets, we show a trend in performance metrics, Error (%), and Ram-hours for different values of <i>bin-width</i> . The value of <i>bin-width</i> varies in [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5]. a) The value of error is evaluated across all the samples for a given value of <i>bin-width</i> . The values of <i>bin-width</i> on x-axis are on logarithmic scale. b) Ram-hours are calculated across all the samples for all the estimators for a given configuration. The values of <i>bin-width</i> on x-axis and Ram-hours on y-axis, both are on log scale. . . . .	67
5.1	The data stream is periodic, has no fluctuation in observed values across different time periods and there is no anomaly. . . . .	76
5.2	A periodic data stream with minor fluctuations and no anomaly. . . . .	76
5.3	A periodic time-series with an anomaly. . . . .	76
5.4	The presence of an anomaly in an aperiodic time-series. . . . .	77
5.5	A data stream with a concept drift. . . . .	77

# List of Tables

2.1	Run-time complexities of algorithms is compared. These complexities are presented with respect to number of samples only. . . . .	33
2.2	FiRE is fast. Execution time (in minutes) collected for the four methods with cell counts varying from 1k to ~68k. . . . .	33
3.1	Run-time complexities of algorithms is compared. These complexities are presented with respect to the number of samples only. . . . .	43
3.2	The performance of FiRE.1 [53], FiRE [67], and HBOS [45] is evaluated on 5 different clusters. A cluster consists of datasets with similar distributions of <i>o-score</i> . Different clusters consist of a different count of datasets. For every cluster, the performances have been compared from the context of all evaluation measures and graded using Friedman ranking. The lowest value of Friedman ranking across methods for a given measure is boldfaced. . . . .	50
3.3	Summary of the large datasets . . . . .	52
3.4	The performances of FiRE.1 [53], FiRE [67], and HBOS [45] are compared on 4 large datasets. The values of the following evaluation measures are reported: <i>Adjusted AP</i> , <i>Adjusted P@n</i> , <i>AP</i> , <i>P@n</i> , and <i>ROC AUC</i> . A method with the highest value of evaluation measure for a given dataset is boldfaced. . . . .	53

4.1	Time complexities of algorithms is compared. This table presents the complexity to process the $N$ samples from a stream. The base estimators are as per the default parameters of the corresponding classes in <code>scikit-multiflow</code> package. In the table $N$ represents number of samples, $d$ represents number of dimensions, $L$ represents number of estimators, $C$ represents the number of classes, $w$ is the window size, $k$ is the number of trials coming from Poisson distribution, and $s$ is the oversampling rate. To be noted, these are the simplified estimates of the time complexity. . . . .	63
4.2	Description of datasets. . . . .	65
4.3	Error (in %) is reported to compare the performance of Enhash [66] with other methods. For a given dataset, the method with the least error is in boldface. Due to implementation constraint, <code>Learn<sup>++</sup></code> .NSE could not run for the <code>outdoorStream</code> dataset. . . . .	68
4.4	KappaM is tabulated to compare the performances of the methods. For a given dataset, the method with the highest value of KappaM is in boldface.	68
4.5	KappaT is reported to compare the performances of the methods. The highest value of KappaT in each row is highlighted. . . . .	69
4.6	The running time of different methods is compared using Time (in hrs). The method with the fastest speed is highlighted for every dataset. . . .	69
4.7	The memory consumption is measured in terms of RAM-hours. The method with the least value of RAM-hours is highlighted for every dataset.	69
4.8	Ablation study of Enhash. The performance of Enhash is compared with its two different variants- 1. Enhash with $\lambda = 0$ (referred to as Enhash-lambda0), and 2. Enhash when ties in <i>concept class</i> assignments are not broken by considering the distance of an incoming sample from the mean of classes in the bucket (referred to as Enhash-noWeights). . . . .	71

# List of Abbreviations

*P@n* Precision at  $n$

*kNN*  $k^{\text{th}}$  Nearest Neighbor

*kNNW* *kNN*-weight

*AP* Average Precision

*AUC* Area Under the Curve

*FPR* False Positive Rate

*ROC AUC* Receiver Operator Characteristics Area Under the Curve

*ROC* Receiver Operator Characteristics

*TPR* True Positive Rate

**ABOF** Angle-based outlier factor

**ADWIN** ADaptive WINdowing

**AEE** Additive Expert Ensemble

**ARF** Adaptive Random Forest

**AWE** Accuracy-Weighted Ensemble

**COF** Connectivity-based Outlier Factor

**CTCs** Circulating Tumor Cells

**DB-outlier score** Distance-based outlier score

**DC** Dendritic Cell

**DCs** Dendritic Cells

**DE** differentially expressed

**EPCs** Endothelial Progenitor Cells

**ESC** embryonic stem cell

**ESCs** embryonic stem cells

**FACS** Fluorescence-activated cell sorting

**FastABOD** Fast Angle-Based Outlier Detection

**FiRE** Finder of Rare Identities

**FP** False Positives

**HBOS** Histogram-based Outlier Score

**INFLO** INFLuenced Outlierness

**KDEOS** Kernel Density Estimation Outlier Scores

**LB** Leveraging Bagging

**LDE** Local Density Estimate

**LDF** Local Density Factor

**LDOF** Local Distance-based Outlier Factor

**LIC** Local Isolation Coefficient

**LOF** Local Outlier Factor

**LoOP** Local Outlier Probabilities

**LSH** Locality Sensitive Hashing

**OB** Online Bagging-ADWIN

**ODIN** Outlier detection using Indegree Number

**OSMOTEB** Online SMOTE Bagging

**PBMC** peripheral blood mononuclear cell

**PBMCs** peripheral blood mononuclear cells

**RMS** Root Mean Square

**scRNA-seq** single-cell RNA sequencing

**SNV** Single Nucleotide Variant

**SOD** Subspace Outlier Degree

**t-SNE** t-Distributed Stochastic Neighbor Embedding

**TN** True Negatives