A SLIDING ADAPTIVE BETA DISTRIBUTION MODEL FOR CONCEPT DRIFT DETECTION IN A DYNAMIC ENVIRONMENT

ANGBERA ATURE

UNIVERSITI SAINS MALAYSIA

2025

A SLIDING ADAPTIVE BETA DISTRIBUTION MODEL FOR CONCEPT DRIFT DETECTION IN A DYNAMIC ENVIRONMENT

by

ANGBERA ATURE

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

June 2025

DEDICATION

I dedicate this work first and foremost to Almighty God, whose boundless grace, wisdom, and strength have guided me through every step of this journey.

To the cherished memory of my beloved daughter, Liana-Ter Godslove Angbera — though your time with us was brief, your presence left an eternal imprint on my heart. This work is a tribute to your beautiful soul.

To my late grandfather, Pa. Angbera Ya Agan, whose legacy of wisdom, integrity, and unwavering faith continues to inspire me.

May this dedication honor your memories and reflect the purpose and strength I have drawn from each of you.

ACKNOWLEDGEMENT

I express my heartfelt gratitude to God Almighty for His unwavering guidance, strength, and blessings throughout this academic journey. My sincere appreciation goes to my supervisor, Associate Professor Dr. Chan H. Y., for his invaluable advice, expertise, and encouragement, which shaped the quality and direction of this work. I am grateful to the Dean and Lecturers of the School of Computer Sciences, Universiti Sains Malaysia, for fostering an environment conducive to learning and research. Special thanks to Joseph Sarwuan Tarka University, Makurdi, for the opportunity and institutional support, and to the Tertiary Education Trust Fund (TETFUND) for their generous sponsorship that made this study possible. To my beloved wife, Mrs. Msendoo Comfort Ature, your love, patience, and sacrifices have been my strength and motivation. I am also thankful to my children, Surun-Ter Praise, Aondosoo Joshua, and Averendoo Israel Angbera for their patience, understanding, and joy that kept me focused. I deeply appreciate the love and support of my parents, Mr. & Mrs. Jacob Ukor Angbera, whose guidance shaped my character. To my siblings, Barnabas, Member, Nguseer Ameen, Doom, Kumawuese, and Sengohol Angbera, thank you for your encouragement and inspiration. My cousin, Msuega Angbera, has been a steady source of support, and I am truly grateful. Special thanks to Prof. Esiefarienrhe, M. B., Prof. Ikughur J. A., and Dr. Ashezua T. T. for their mentorship and encouragement, which greatly influenced my academic and professional development. I also acknowledge Rev. & Atese Terkaa, J. T. for their spiritual guidance and prayers. Lastly, I appreciate my friends Godwin Efenji, Sabastine Emmanuel, and many others for their encouragement, laughter, and belief in me. Your support made this journey meaningful and memorable.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ADWIN	Adaptive Sliding Window
AFXGB	Adaptive Fast XGBoost
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
ARF	Adaptive Random Forest
AUC	Area Under the Curve
AUE	Accuracy Updated Ensemble
AWE	Accuracy Weighted Ensemble
BDDD	Beta Distribution Drift Detection
BI	Business Intelligence
CER	Classification Error Rate
CPs	Change Points
CPU	Central Processing Unit
CUSUM	Cumulative Sum
DA	Data Analytics
DDM	Drift Detection Method
DDM-OCI	Drift Detection Method Online Class Imbalance
DMWM	Dynamic Modeling With Memory
DW-CAV	Dynamically Weighted Consult and Vote
DWM	Dynamic Weighted Majority
EDDM	Early Drift Detection Method
EF	Ensemble-based Framework
EFDT	Extremely Fast Decision Tree
FN	False Negative

FNR	False Negative Rate
FNR	False Negative Rates
FP	False Positive
FPDD	Fisher Proportion Drift Detector
FPR	False Positive Rate
FPR	False Positive Rates
FSDD	Fisher-based Statistical Drift Detector
FTDD	Fisher Test Drift Detector
GP	Gaussian Process
HDFS	Hadoop Distributed File System
HPO	Hyperparameter Optimization
HT	Hoeffding Tree
HTM	Hierarchical Temporal Memory
IOT	Internet of Things
KL	Kullback-Leibler Divergence
KNN	K-Nearest Neighbors
LB	Leveraging Bagging
LCDD	Local Complete-based Drift Detection
LFR	Linear Four Rates
MDDMs	McDiarmid Drift Detection Methods
MDDT	Multiscale Drift Detection Test
MET	Multiple Event Types
ML	Machine Learning
MOA	Massive Online Analysis
MSE	Mean Square Error
OCDD	One Class Drift Detector
OCI-CD	Online Class Imbalance and Concept Drift

OPA	Online Passive-Aggressive
OS-ELMs	Online Sequential Extreme Learning Machines
PAC	Probably Approximately Correct
PH	Page Hinckley
PT	Prediction Tools
RDD	Resilient Distributed Dataset
RDDM	Reactive Drift Detection Method
SABeDM	Sliding Adaptive Beta Distribution Model
SAM-KNN	Self-Adjusting Memory KNN
SAND	Semi-supervised Adaptive Novel-class Detection
SD	Standard Deviation
SEA	Streaming Ensemble Algorithm
SI	Statistical Inference
SLT	Statistical Learning Theory
SMOTE	Synthetic Minority Over-sampling Technique
SPRT	Sequential Probability Ratio Test
SRP	Streaming Random Patches
STEPD	Statistical Test of Equal Proportion
TN	True Negative
TP	True Positive
TPE	Tree Parzen Estimator
TPR	True Positive Rate
XGBoost	eXtreme Gradient Boosting

LIST OF APPENDICES

MODEL AGIHAN BETA ADAPTIF GELANGSAR UNTUK PENGESANAN HANYUTAN KONSEP DALAM PERSEKITARAN DINAMIK

ABSTRAK

Model pembelajaran mesin yang digunakan dalam persekitaran aliran data sering mengalami perubahan konsep, iaitu perubahan dalam taburan data dari semasa ke semasa yang menyebabkan penurunan prestasi. Mengesan dan menyesuaikan diri dengan perubahan ini secara masa nyata adalah sangat penting untuk mengekalkan ketepatan dan kebolehpercayaan model. Kajian ini memperkenalkan Model Agihan Beta Adaptif Gelangsar (SABeDM), satu pendekatan inovatif untuk pengesanan dan penyesuaian terhadap perubahan konsep dalam aliran data yang dinamik. SABeDM memanfaatkan mekanisme pengesanan perubahan berdasarkan agihan beta untuk mengenal pasti perubahan dalam agihan data dan mencetuskan kemas kini model secara automatik. Teknik tetingkap gelangsar digunakan untuk memastikan kemas kini masa nyata apabila data baharu tersedia. Selain itu, model ini menggabungkan faktor peluruhan masa untuk meningkatkan kebolehsuaian dengan memberi keutamaan kepada data terkini serta menggunakan Divergensi Kullback-Leibler (KL) dan mekanisme toleransi masa bagi mengurangkan kesilapan pengesanan (positif palsu dan negatif palsu) dalam mengenal pasti perubahan konsep. Untuk menangani cabaran dalam senario perubahan berkala, di mana kaedah tradisional menghadapi kesukaran menangani saiz keratan yang melebihi tetingkap pemprosesan, SABeDM memperkenalkan mekanisme tetingkap adaptif yang bertindih serta strategi pembelajaran masa adaptif, sekali gus meningkatkan keberkesanan pengesanan perubahan. Hasil eksperimen ke atas set data sintetik dan dunia sebenar menunjukkan bahawa SABeDM mengatasi kaedah sedia ada. Model ini berjaya mencapai

peningkatan ketepatan pengesanan perubahan sebanyak +12.89% pada SEA_g, pengurangan positif palsu/negatif palsu sebanyak +13.74% untuk ketepatan SEA_g, serta peningkatan prestasi klasifikasi penyesuaian selepas hanyut perubahan sebanyak +12.78% untuk ingatan semula SEA_g. Peningkatan lain termasuk +4.97% peningkatan ketepatan untuk ARG_a dan +5.28% untuk ARG_g, +5.94% peningkatan ketepatan untuk ARG_a dan +6.77% untuk ARG_g, +3.78% peningkatan skor F1 untuk MIXD, +3.51% peningkatan ketepatan untuk WET, serta +3.06% peningkatan ketepatan untuk PHI, dengan peningkatan ketepatan dan ingatan semula masing-masing sebanyak +2.49% dan +2.70%. Keputusan ini menunjukkan keberkesanan SABeDM dalam mengesan dan menyesuaikan diri terhadap perubahan konsep, serta menawarkan satu penyelesaian yang menjanjikan untuk pembelajaran adaptif dalam persekitaran data yang dinamik.

A SLIDING ADAPTIVE BETA DISTRIBUTION MODEL FOR CONCEPT DRIFT DETECTION IN A DYNAMIC ENVIRONMENT

ABSTRACT

Machine learning models deployed in data streaming environments often suffer from concept drift, where the underlying data distribution changes over time, leading to performance degradation. Detecting and adapting to these shifts in real time is crucial to maintaining model accuracy and reliability. This study introduces the Sliding Adaptive Beta Distribution Model (SABeDM), a novel approach for concept drift detection and adaptation in dynamic data streams. SABeDM leverages a beta distribution-based drift detection mechanism to identify distributional changes and trigger model updates accordingly. A sliding window technique is incorporated to ensure real-time updates as new data becomes available. The model also integrates a time decay factor, enhancing its adaptability by prioritizing recent data, and employs Kullback-Leibler (KL) Divergence and a time tolerance mechanism to reduce false positives and false negatives in drift detection. To address challenges in periodic drift scenarios where traditional methods struggle with chunk sizes exceeding processing windows, SABeDM introduces an overlapping adaptive window mechanism and an adaptive learning time strategy, ensuring more effective drift detection. Experimental results on synthetic and real-world datasets demonstrate that SABeDM outperforms existing methods, achieving: +12.89% increase in drift detection accuracy on SEA_g, +13.74% reduction in false positives/negatives for SEA_g precision, +12.78% improvement in classification performance post-drift adaptation for SEA_g recall, +4.97% improvement in accuracy for ARG_a and +5.28% for ARG_g, +5.94% increase in precision for ARG_a and +6.77% for ARG_g, +3.78% increase in F1-score

for MIXD, +3.51% improvement in accuracy for WET, +3.06% increase in accuracy for PHI, with precision and recall improvements of +2.49% and +2.70%, respectively. SABeDM proves effective in concept drift detection, offering a promising solution for adaptive learning in dynamic environments.

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Artificial intelligence (AI) is the field that focuses on creating computer systems with artificial intelligence, a concept debated by scholars over the years (Entwistle, 1988; Hamet and Tremblay, 2017; Mintz and Brodie, 2019). The past decade has seen a significant surge in AI interest, with machine learning (ML) at its core gaining widespread popularity (Batta, 2018; Carbonell, 1981; Hua and Learning, 2009; Jordan and Mitchell, 2015). ML, which involves learning from extensive historical data, facilitates tasks like prediction, classification, and recognition. Companies are actively exploring advanced ML techniques for a competitive edge, impacting fields from autonomous driving (Janai et al., 2020) to natural language processing applications (Brown et al., 2020). Industries also benefit from ML-based preventive maintenance to minimize production downtimes (Carvalho et al., 2019; Ersöz et al., 2022), while real-time models or applications grapple with the challenge of Concept Drift, where data distribution evolves.

Concept drift is a problem that directly affects the performance of ML classification systems, as these systems tend to become less effective over time (Lu et al., 2019), meaning that high recognition rates may not be attained. Today, supervised ML is the most used ML methodology. This method aims to learn a function that translates the input data X to a corresponding label y (Jordan and Mitchell, 2015). A model is realised by considering a series of training data samples where input (X) and label (y) information is available. Once a model has been trained, it may be used to predict fresh data examples. However, in a dynamic environment or real-time learning,

the link between the input data X and the target y, or the underlying distribution of the input data X, could alter with time (Gama et al., 2013). Concept drift (Widmer and Kubat, 1996) is the name for this problem. Concept drifts can significantly affect the underlying prediction model's predictive accuracy (Lindstrom et al., 2013; Yang et al., 2021b; Yang and Shami, 2022; Zliobaite et al., 2016a).

Anti-spam filters, weather forecasts, surveillance systems, fraud detection, customer preferences, crime forecasts, disease forecasts, and environmental monitoring are a few issues with dynamic environments. For instance, the characteristics of spam can change over time when using anti-spam filtering. Key features currently utilised to define spam may also become obsolete (Kuncheva, 2004a). Therefore, the anti-spam filter needs a technique to detect changes to adjust to new spamming patterns. Similarly, ML models in climate forecasting must continuously adapt to shifting weather patterns to provide accurate predictions (Khashei and Bijari, 2010). In e-commerce, recommender systems must detect changes in consumer preferences over time to provide relevant product recommendations (Al-Muhaideb et al., 2022).

Proper solutions must be used to address this issue because ML models significantly impact many vital organizational components. Without this, machine learning-optimized complex systems used today (such as multi-stage supply chain networks) would be unable to deliver on their value proposition, which will have significant repercussions. For instance, studies have shown that concept drift in financial fraud detection can lead to an increase in undetected fraudulent transactions (Dal Pozzolo et al., 2017), while in healthcare, outdated predictive models may compromise patient outcomes (Tsymbal, 2004; Webb et al., 2018). In cybersecurity,

attackers frequently adapt their strategies, requiring continuous adaptation of ML-based intrusion detection systems to remain effective (Mala et al., 2021; Khammas et al., 2023).

Additionally, in social media content moderation, evolving user behavior and linguistic trends necessitate adaptive filtering techniques to ensure accurate and fair content moderation (Zhang et al., 2022). Furthermore, in e-commerce, changes in consumer behavior and seasonal trends can lead to outdated recommendation models, necessitating the continuous adaptation of personalization algorithms (Al-Muhaideb et al., 2022). Likewise, in autonomous systems such as self-driving cars, changes in environmental conditions and road patterns require adaptive learning techniques to ensure safe and reliable operation (Bansal et al., 2023).

The literature is replete with studies that suggest innovative ways to create classification systems with machine learning models (Bayram et al., 2022; Priya and Uthra, 2020; Raab et al., 2020; Shahraki et al., 2022; Yan, 2020) that can recognise changes and update their knowledge without affecting the system's accuracy. Many approaches, however, have concentrated on either retraining the classifier without explicitly identifying changes or detecting changes based on tracking the system's success rate. In the first scenario, a sudden decrease in system performance is required to identify changes, which unquestionably indicates system failure. Since the system updates continuously, even when changes do not occur, the fundamental drawback of this category of techniques is the computational expense involved. In addition, when the system relies solely on a single classifier, the old concept may be forgotten as soon as the new one is absorbed. Due to such behaviour, a system could experience

catastrophic forgetting (Ebrahimi et al., 2021; French, 1993; Ramasesh et al., 2022; Shmelkov et al., 2017).

In today's data-driven world, the constant evolution and transformation of data distributions have become a pivotal challenge across various fields, including ML, data mining, and AI. This phenomenon, known as concept drift, as stated earlier, occurs when the statistical properties of data change over time, rendering traditional static models and algorithms inadequate. In response to this challenge, researchers and practitioners have been actively seeking innovative methods and techniques to detect and adapt to concept drift in dynamic data environments.

Concept drift detection is a critical aspect of many real-world applications, as the stability and accuracy of predictive models are often undermined by the changing nature of data distributions. The consequences of failing to address concept drift can be severe, leading to decreased performance, incorrect predictions, or even costly errors. Real-world applications such as fraud detection, network security, recommendation systems, and predictive maintenance rely heavily on the ability to detect and adapt to concept drift in a timely and accurate manner. This research has highlighted that integrating adaptive ML strategies, such as adaptive sliding window and incremental model updates, can significantly mitigate the adverse effects of concept drift (Gama et al., 2014; Krawczyk et al., 2017; Žliobaitė et al., 2016b). Therefore, addressing concept drift remains a fundamental challenge in the AI and ML communities, necessitating continuous research and innovation.

1.2 Research Motivation

This thesis is motivated by the pressing need to address the challenges faced by ML models in environments where data distributions continuously evolve. Concept

drift a phenomenon where data patterns shift over time poses a significant challenge to ML models operating in dynamic settings such as financial markets, healthcare, and social media analytics. In financial markets, for example, trading patterns and economic conditions fluctuate, requiring ML models to adapt to new trends to avoid inaccurate predictions. In healthcare, disease progression and treatment responses vary over time, making it crucial for diagnostic models to detect and adjust to these changes to provide accurate recommendations.

Similarly, in social media analytics, user behaviors and content trends evolve rapidly, demanding adaptive models that can capture emerging patterns for targeted recommendations and sentiment analysis. If left unaddressed, such changes can lead to performance degradation, rendering models ineffective or even misleading. Therefore, the primary motivation for this research stems from the recognition of concept drift as a pervasive issue that must be tackled to ensure the reliability and effectiveness of ML models in real-world applications. This is because the use of ML enhances data management and decision support systems. While this technology is currently in its early stages (Baier, 2021), it can eventually improve every single decision or information system component (Bawack et al., 2019; Senthil Kumar, 2020; Sutton et al., 2020).

The significance of concept drift detection becomes evident in domains where decision-making relies on up-to-date, accurate predictions. For instance, in fraud detection systems, financial transactions exhibit evolving patterns due to changes in fraud strategies (Bian et al., 2022; Crespo and Weber, 2005; Gokasar et al., 2023; S and Almutairi, 2023). A failure to detect and adapt to such shifts could result in undetected fraudulent activities, leading to significant financial losses. Similarly, in

autonomous vehicles, environmental conditions such as lighting, traffic patterns, and pedestrian behaviors change dynamically (Idiri and Napoli, 2012; Li et al., 2023; L. Ma et al., 2022). An ML model incapable of adapting to such changes may compromise safety. These examples illustrate why developing an effective concept drift detection method is crucial for ensuring the continued effectiveness of ML models in rapidly changing environments.

Additionally, managing concept drift effectively aligns with the United Nations' Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-Being), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 11 (Sustainable Cities and Communities) (United Nations, 2015). In healthcare applications, timely adaptation to new disease patterns or health trends can improve medical diagnostics and treatment plans, ultimately contributing to better health outcomes (Dargan and Kumar, 2020; Fenu and Marras, 2021; Sharma and Ross, 2023). In smart cities, adaptive ML systems for energy management, traffic monitoring, and public safety can foster more sustainable and efficient urban environments (Anguita, 2001; Kaushik et al., 2022; Yang et al., 2021). This connection highlights how addressing drift challenges not only improves model performance but also contributes to broader societal goals.

Another key motivation is the efficient management of computational resources in ML systems. An adaptive approach for concept drift detection, where models continuously update without verifying whether drift has actually occurred, can lead to unnecessary computational costs and potential overfitting to noise. This issue, known as blind adaptation, results in wasted resources and degraded model performance. Thus, a strong motivation for this research is the need to develop drift

detection techniques that intelligently discern when updates are necessary, ensuring computational efficiency while maintaining model accuracy. Resource-efficient methods are especially important in sustainable computing efforts, supporting SDG 12 (Responsible Consumption and Production) by minimizing energy consumption and optimizing resource utilization (United Nations, 2015).

As real-world applications increasingly demand adaptive learning capabilities, the motivation for effective concept drift detection extends across multiple industries. In forestry, shifts in climate conditions and land use require models that can track and respond to ecological changes (Sanquetta et al., 2013; Singh et al., 2021; Wood, 2021). Telecommunications networks must detect shifts in traffic patterns to optimize service delivery (Hilas, 2009; Mazhelis and Puuronen, 2007), while security systems must adapt to evolving cyber threats (Mao et al., 2018; Sadoddin and Ghorbani, 2008, 2009). Other fields, including transportation (Bian et al., 2022; Crespo and Weber, 2005; Gokasar et al., 2023; S and Almutairi, 2023), maritime systems (Idiri and Napoli, 2012; Li et al., 2023; L. Ma et al., 2022), and online recommendation systems, also depend on ML models that can dynamically adjust to changing data patterns, the motivation becomes clear.

Given the widespread impact of concept drift, the motivation behind this research lies in developing effective, efficient, and adaptive ML techniques. By enabling ML models to detect and respond to drift effectively, this research contributes to enhancing their long-term reliability and decision-making accuracy in a constantly evolving data landscape. Additionally, these efforts align with global initiatives to leverage technological advancements in building more sustainable and resilient societies in line with the SDGs (United Nations, 2015).

1.3 Problem Statement

Machine learning (ML) models deployed in dynamic environments face the challenge of concept drift (Suárez-Cetrulo et al., 2023; Wares et al., 2019), where changes in data distributions over time lead to model degradation if not properly detected and addressed (Adam et al., 2020; Zhang et al., 2022). Studies have shown that failing to detect and adapt to concept drift can result in accuracy degradation of up to 30% in real-world applications such as fraud detection, predictive maintenance, and healthcare analytics (Gama et al., 2014; Žliobaitė, 2010; Lu et al., 2018; Olowu et al., 2024). Effective concept drift detection is essential to maintaining model reliability (Himaja et al., 2019; Pears et al., 2014; Yan, 2020), but existing techniques suffer from performance limitations in balancing detection accuracy and computational efficiency (Hovakimyan and Bravo, 2024).

A commonly used approach, blind or passive adaptation, updates models at regular intervals without verifying whether a drift has actually occurred (Amutha et al., 2020; Song et al., 2022, 2019). While this ensures continuous learning, it often results in unnecessary retraining, leading to excessive computational costs and potential overfitting to transient noise rather than genuine data shifts (Widmer & Kubat, 1996; Gama et al., 2014). This predefines intervals for regular update usually makes detection of other patterns of drift difficult to detect. Techniques such as statistical hypothesis testing, window-based methods, and ensemble approaches have been proposed to mitigate these issues, yet each comes with trade-offs in detection delay, computational complexity, and adaptability to different drift types, including sudden, gradual, incremental, and recurring drifts. Achieving an optimal balance between these errors (false positives and false negatives) is difficult for traditional methods (Baena-García et al., 2006; Bifet & Gavaldà, 2007; Huang et al., 2015).

High false positive rates in drift detection lead to redundant model updates (Ali and Mahmood, 2024; Pesaranghader et al., 2018), wasting computational resources and increasing training overhead, whereas high false negative rates allow outdated models to persist, reducing classification accuracy and decision-making reliability (Agrahari and Singh, 2022; Albert Bifet and Gavaldà, 2007; João Gama et al., 2004; Halder and Hasan, 2022; Huang et al., 2015). Furthermore, according to the probably approximately correct (PAC) learning model (Mitchell, 1997; Valiant, 1984), a significant false positive rate would prevent accuracy from improving because a negligible quantity of data would be used for training. Streaming data is dynamic; therefore, the drift detection algorithms shouldn't assume the incoming input data, like prediction results, follow a particular distribution function (Frías-Blanco et al., 2015; Xuan et al., 2020). Existing drift detection approaches often assume specific data distributions, making them less effective in real-world streaming applications where data characteristics are unknown and continuously evolving (Ditzler et al., 2015).

Given these challenges, there is a critical need for adaptive drift detection mechanisms that dynamically assess the necessity for model updates. Instead of updating models at predefined intervals, an ideal system should leverage advanced statistical and probabilistic measures to accurately determine drift occurrences. Methods integrating change-point detection, adaptive windowing, and information-theoretic metrics have shown promise (Agrahari and Singh, 2025; Ogasawara et al., 2025; Du et al., 2014), but challenges remain in balancing detection sensitivity with computational efficiency (Gu et al., 2024; Wan et al., 2024).

Despite extensive research on concept drift detection, existing methods struggle to achieve a balance between detection accuracy and computational efficiency

in dynamic environments. Many approaches either rely on passive adaptation, which leads to unnecessary retraining and resource wastage, or employ statistical and ensemble-based techniques that suffer from trade-offs in sensitivity, delay, and adaptability to different drift types. These limitations highlight the need for an adaptive, efficient, and scalable drift detection model that minimizes false detections, optimizes model adaptation strategies, and ensures reliable performance in dynamic data environments. Therefore, this research seeks to develop novel concept drift detection model that enhance detection accuracy and optimize model adaptation strategies. By leveraging advanced statistical techniques and adaptive learning frameworks, the proposed solution aims to surpass existing methodologies in drift detection rate, and classification performance in dynamic environments.

1.4 Research Questions

The significance of this study will be encapsulated through the exploration of three fundamental questions.

1. How can a probabilistic sliding window approach be leveraged to enhance the detection of evolving patterns in dynamic data streams?
2. What techniques can improve the adaptability of concept drift detection models while preventing unnecessary updates and overfitting?
3. How can optimization strategies be applied to achieve a balance between sensitivity and accuracy in detecting genuine distributional changes?

1.5 Research Objectives

The main aim of this research is to suggest an approach based on error monitoring to deal with data streams by explicitly detecting drifts and responding to

them without degrading classification performance while achieving comparable or higher accuracy and detection rates compared to existing solutions. The objectives of this research are:

1. To develop a novel concept drift detection approach that leverages the Beta probability distribution and adaptive sliding windows to improve classification accuracy in dynamic data environments.
2. To enhance the adaptability of the proposed model by incorporating adaptive sliding window mechanisms with a time decay factor to minimize blind adaptation and maintain higher classification accuracy.
3. To optimize the proposed approach using a time tolerance mechanism and Kullback-Leibler (KL) divergence, effectively distinguishing between noise and genuine drift, thereby reducing false positive and false negative rates compared to existing models.

1.6 Research Significance

Concept drift significantly impacts the stability and predictive accuracy of machine learning (ML) models in dynamic environments. Undetected drift leads to model degradation, increased false predictions, and unreliable decision-making in real-time data streams. This research introduced the Sliding Adaptive Beta Distribution Model (SABeDM), which integrates Beta probability distribution modeling within an adaptive sliding window framework to enhance drift detection accuracy and model adaptability.

The study contributes to concept drift detection by employing Beta probability distribution modeling to estimate changes in data distributions. This statistical approach allows for the differentiation between genuine drift and random fluctuations,

improving the reliability of detection. Unlike conventional methods that rely on heuristic-based thresholds such as DDM (Drift Detection Method) (Gama et al., 2004), EDDM (Early Drift Detection Method) (Baena-García et al., 2006), and HDDM (Hoeffding's Bounds Drift Detection Method) (Frias-Blanco et al., 2015) which set predefined confidence intervals for drift detection, the probabilistic nature of the Beta distribution provides a more systematic and data-driven approach to identifying shifts in data distribution.

The proposed SABeDM is designed to minimize both false positives and false negatives in concept drift detection, ensuring efficient adaptation to dynamic data environments. By integrating Beta distribution modeling, adaptive sliding windows, and Kullback-Leibler (KL) divergence, SABeDM reduces false positives by preventing unnecessary model updates triggered by minor fluctuations or noise. Simultaneously, it lowers false negatives by accurately identifying significant distribution changes, preventing outdated models from making incorrect predictions. This balance enhances classification accuracy, optimizes computational resources, and improves the overall reliability of machine learning models in real-time streaming data applications.

Adaptability to evolving data patterns is another crucial aspect of drift detection. Fixed-size sliding windows used in traditional models limit the responsiveness of drift detection mechanisms, as they fail to adjust dynamically to varying data distributions. To overcome this limitation, the study incorporates an adaptive sliding window mechanism with a time decay factor. This allows the window size to adjust based on the rate of change in the data distribution, preventing

unnecessary updates due to transient variations while ensuring prompt adaptation to significant drifts.

Handling varying chunk sizes in data streams is also a major challenge for existing concept drift detection methods. Many conventional approaches struggle when the chunk size exceeds the processing window, leading to incomplete analysis and inaccurate drift detection. For instance, ADWIN (Adaptive Windowing) (Bifet and Gavaldà, 2007) dynamically adjusts its window size but can still suffer from delays in drift detection when chunk sizes fluctuate significantly. Similarly, Page-Hinkley Test (PHT) (Page, 1954) relies on cumulative error monitoring but may fail to promptly detect drifts when data chunks are irregular. HDDM (Hoeffding's Bound Drift Detection Method) (Frias-Blanco et al., 2015) also encounters difficulties in balancing sensitivity across different chunk sizes, often resulting in excessive false positives or false negatives. To address these limitations, the proposed SABeDM model introduces an overlapping mechanism within the adaptive sliding window and an increased learning time mechanism, ensuring effective drift detection regardless of chunk size variations. This approach allows for continuous monitoring and adaptation, improving accuracy in dynamic data stream environments.

The effectiveness of the proposed model is validated using both synthetic and real-world datasets across multiple application domains, including fraud detection, and recommender systems. The results demonstrate that SABeDM consistently outperforms existing concept drift detection models in terms of accuracy, adaptability, and computational efficiency.

Furthermore, the study aligns with the United Nations Sustainable Development Goals (SDGs) by improving AI-driven decision-making in critical

sectors. Specifically, it contributes to SDG 3 (Good Health & Well-being) by enhancing machine learning models for real-time patient monitoring and diagnostics (United Nations, 2015). It also supports SDG 9 (Industry, Innovation & Infrastructure) by improving the efficiency of industrial automation through adaptive learning mechanisms. Additionally, it advances SDG 11 (Sustainable Cities & Communities) by enabling smart city applications, including intelligent traffic monitoring and real-time security systems.

1.7 Research Scope

This thesis's primary focus is classification and adaptability in the face of changing streaming data. The data streams employed in this research are essentially cross-sectional data and changing datasets over time. It is expected that the data streams have previously been labelled. Future work will likely focus on solving the problems caused by unlabelled instances and delayed labels.

The literature (Abbasi et al., 2021; Agrahari and Singh, 2021; Gama et al., 2004; Guo et al., 2023; Ud Din et al., 2020; Wang et al., 2022; Yang et al., 2021a) frequently utilizes both synthetic and real-world data streams, which are also employed in this research. These data streams contain numerical or categorical properties, and concept drift in synthetic data streams may be develop either suddenly or gradually. To ensure the effectiveness of the proposed approaches, noise (10%) was introduced into some of the synthetic data streams, and drift detectors were evaluated for their ability to distinguish between concept drift and noise (Guo et al., 2023; Krempl et al., 2014; Wang et al., 2022).

The synthetic datasets used include Agrawal (financial domain), SEA (IoT sensor data), MixedDrift, Hyper Plane (robotics), and SINE, which are widely used benchmarks for testing concept drift detection methods due to their controlled drift characteristics and varying degrees of complexity. These datasets allow for rigorous testing of the model's adaptability to sudden, gradual, incremental, recurring drifts under different scenarios. Additionally, the real-world datasets used in the study are Phishing (cybersecurity domain) and Weather (meteorology domain) datasets, which provide practical insights into how the proposed model performs in real-world applications where drifts occur naturally over time. Only datasets with a binary class were used in this research to ensure consistency in classification tasks and evaluation. By incorporating both synthetic and real-world datasets, the study ensures a thorough evaluation of the model's effectiveness, adaptability, and effectiveness in handling concept drift in dynamic data streams.

The study opts for real concept drift due to its numerous benefits, including real-world relevance, enhanced model effectiveness, data-driven insights, evaluation validity, and practical applicability. These advantages underscore the preference for real concept drift over virtual concept drift in research.

To evaluate the drift detection approach for adaptive learning, metrics such as classification accuracy, true positive rate, false positive and negative rates, and drift detection delays are considered. However, the study faces limitations in several areas, including computational complexity, sensitivity to parameter tuning, the necessity of batch processing, and challenges related to unlabeled instances and delayed labels. These limitations highlight the need for future research to develop solutions that address these critical issues.

1.8 Research Contributions

This thesis presents a substantial innovation in the realm of ML and data mining by introducing an improved concept drift detection technique. The ensuing list outlines the noteworthy contributions achieved through this research:

Development of the Sliding Adaptive Beta Distribution Model (SABeDM): The primary contribution of this research is the development of the Sliding Adaptive Beta Distribution Model (SABeDM) for concept drift detection in dynamic data environments. The proposed model leverages the Beta probability distribution to quantify distributional changes and detect drift events effectively. It integrates an adaptive sliding window mechanism that dynamically adjusts to evolving data distributions, thereby improving classification accuracy. To prevent unnecessary retraining, the model incorporates an error monitoring and statistical thresholding approach that triggers updates only when significant drift is detected. This ensures that the classifier remains accurate while reducing computational overhead associated with frequent model updates.

Integration of Adaptive Sliding Windows with Time Decay Factor: To enhance adaptability, the research introduced an adaptive sliding window mechanism with a time decay factor that prioritizes recent data while gradually diminishing the influence of older observations. This approach mitigates blind adaptation, which occurs when a model updates unnecessarily in response to minor fluctuations, leading to instability in classification performance. By dynamically adjusting the adaptation process, the proposed approach ensures a more stable and reliable response to concept drift.

Optimization of Drift Detection Using Time Tolerance and KL Divergence: Further optimization is achieved through the integration of a time tolerance mechanism and Kullback-Leibler (KL) Divergence for drift detection. The time tolerance

mechanism helps distinguish between temporary fluctuations and genuine concept drift, reducing false positive detections that could lead to redundant model retraining. Additionally, KL Divergence measures the statistical distance between data distributions before and after a detected drift, ensuring that only significant distributional changes trigger adaptation. These enhancements lead to a substantial reduction in false positive and false negative rates, improving the overall reliability and efficiency of drift detection.

1.9 Thesis Structure

In this chapter, the background of concept drift has been presented, along with an explanation of the research problem statement, the motivation, research questions, objectives, significance, scope, and contributions. The remaining chapters of this thesis are organized as follows:

Chapter 2 covers pertinent linked literature. An overview of studies that take concept drift and handling methods into account and ML in the context of data streams follows. The fundamental elements of concept drift detection and adaptation were also covered in this chapter. A categorization of the available algorithms based on implementation details was also provided. The final discussion focuses on the algorithms' shortcomings and serves as motivation for the remaining chapters and remedies.

Chapter 3 suggests the mechanism for detecting drift that treats anomalous behaviour as drift and fits a beta distribution to the model error in an adaptive sliding window. Extensive tests are run on synthetic and real-world data streams to compare the new methodology with the state-of-the-art. Experimental results of the proposed

solution for handling concept drift in data streams in dynamic environments are presented in the next chapter.

Chapter 4 presents the proposed model for handling concept drift in dynamic environments using ML models for classification. Chapter 5 presents the results and discussions of all the experiments carried out in this research from the proposed concept drift handling methods in dynamic environments using ML models for classification. The conclusions of this thesis are outlined in Chapter 6, along with suggestions for further research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter establishes the theoretical foundation essential for understanding the contributions of this thesis. It introduces fundamental concepts, key definitions, and relevant works from related fields. By positioning the research within existing literature, this section provides a foundational grasp of these strategies, ensuring a smoother transition into the more technical discussions in subsequent chapters.

The chapter concludes by summarizing the key points discussed. It highlights the significance of understanding concept drift in adaptive machine learning and emphasizes the necessity of adaptive techniques in dynamic settings. The groundwork laid in this chapter forms the basis for subsequent explorations in this work.

2.2 Concept Drift

In this segment, fundamental terminology and principles concerning concept drift are delineated. Initially, the characterization of concept drift is expounded upon, followed by a concise survey of predominant strategies for detecting it. The subsequent sections of this chapter elucidate pivotal notions pertinent to investigations into detecting and mitigating concept drift.

2.2.1 Definition of Concept Drift

The phrases "concept" and "concept drift" have distinct meanings concerning ML. A concept is a model that depicts how input and output variables are connected ambiguously and concealed (Balk, 2021). An excellent example is a meteorological data concept, where the season is not explicitly stated in the temperature information

but influences the temperature information. That concept must be captured for a classification model to produce accurate predictions.

Defining a concept mathematically and portraying it as a probabilistic connection is possible. The random variable $\mathcal{Y}$ is used to identify the type of data point or label, whereas $\mathcal{X}$ denotes the random variables present in a set of vectors, or features, assigned to a given data point. For a given data point, the variables $\mathcal{X}$ and $\mathcal{Y}$ are connected, with $\mathcal{P}(\mathcal{X})$ denoting the prior probability distribution over labels and $\mathcal{P}(\mathcal{Y})$ denoting the distribution over covariates (Webb et al., 2016). Joint distribution is used to define a concept, $\mathcal{C}$, as shown in Equation 2.1 (Webb et al., 2016):

$$\mathcal{C} = \mathcal{P}(\mathcal{X}, \mathcal{Y}) \tag{2.1}$$

A good example of the Equation 2.1 can be seen in predicting the weather. Suppose $\mathcal{X}$ represents input features like temperature, humidity, and wind speed, while $\mathcal{Y}$ is the output label, such as "rain" or "no rain." The joint probability $\mathcal{P}(\mathcal{X}, \mathcal{Y})$ describes the likelihood of certain weather conditions leading to rain. For instance, if high humidity and dark clouds often lead to rain, the model learns this pattern as a concept. However, if climate conditions change over time such as increasing temperatures leading to different rainfall patterns the original relationship between $\mathcal{X}$ and $\mathcal{Y}$ no longer holds. This shift in $\mathcal{P}(\mathcal{X}, \mathcal{Y})$ means the model needs to adapt to new weather patterns to make accurate predictions.

Concept drift is the term for a change in a concept that could result in misclassification by a particular predictor. A shift in the distribution where the examples are drawn is the cause of it. Non-stationarity is another name for this. Concept drift happens when the underlying feature data in a dataset undergoes a probabilistic

distributional shift over time. In other words, the relationship between the past and current concepts is no longer valid. The concept has strayed from the earlier connection.

According to Bayesian decision theory, classification can be explained by the prior probabilities of the classes $\mathcal{P}(c)$ and the class conditional probability density function $\mathcal{P}(\mathcal{X}|c)\ \forall\ c\ in\ \mathcal{C}$ (Duda et al., 2000). The posterior probability can be used to determine if a given sample $\mathcal{X}$ belongs to class c as shown in Equation 2.2.

$$\mathcal{P}(c|\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}|c)\ .\mathcal{P}(c)}{\mathcal{P}(\mathcal{X})} \tag{2.2}$$

Accordingly, concept drift is any scenario in which the posterior probability changes with time *t* (Agrahari and Singh, 2022; Elwell and Polikar, 2011). Concepts typically depend on the context of the relevant data stream, which is frequently hidden in the ML model (e.g., important factors that are not included in the input features of the model). Because the ML model is unable to detect changes in the context, making accurate predictions is difficult (Widmer & Kubat, 1996). Customer preferences that change over time are a good illustration of concept drift. A technique for predicting customer purchasing behaviour is used in this situation. One of the customers receives an unexpectedly large salary boost that the algorithm cannot detect (changing hidden context). Owing to her altered income, the consumer has modified her shopping habits, such as by purchasing more organic goods, which makes it challenging to provide the ML model with high-calibre recommendations. Concept drift is described as follows (Gama et al., 2014; Widmer & Kubat, 1996) as shown in Equation 2.3.

$$\mathcal{P}_t(c|\mathcal{X})\ \neq\ \mathcal{P}_{t+1}(c|\mathcal{X}) \tag{2.3}$$

where $\mathcal{P}_t(c|\mathcal{X})$ and $\mathcal{P}_{t+1}(c|\mathcal{X})$ are concepts at various times. Thus, *t* and *t+1* are two distinct points in time, where $t < t+1$.

The ML field uses additional terminology in addition to concept drift to explain related phenomena referring to shifting data distributions (Moreno-Torres et al., 2012). For instance, dataset shift is defined as a change in the shared probability distribution of input data $\mathcal{X}$ and related classes or labels c between training *(trt)* and test *(tst)* time. (Quionero-Candela et al., 2009). This can be mathematically seen in Equation 2.4.

$$\mathcal{P}_{trt}(c|\mathcal{X}) \neq \mathcal{P}_{tst}(c|\mathcal{X}) \quad 2.4$$

The indices are essential for distinguishing between the two definitions: Concept drift is strongly related to the issue of ML in data stream settings since it refers to the temporal element of the data, whereas dataset shift focuses on the distinction between the training and testing environments.

Models are built using ML techniques, and the models are then utilized to assess and comprehend reliable datasets. To predict and classify data points, the models are applied to datasets. The machine-learned classification algorithms are unable to accurately identify points in a dataset when concept drift takes place. The dataset must be used to update the model using the most recent data.

However, because the class prior probabilities $\mathcal{P}(c)$ are independent of the characteristics, witnessing a change in the evidence $\mathcal{P}(\mathcal{X})$ is insufficient to detect a concept drift. The decision limits within the classes are more telling. *Virtual concept drift* is the term used to describe situations when the boundaries stay the same, but the likelihood $\mathcal{P}(\mathcal{X}|c)$ varies. In other words, $\mathcal{P}_t(\mathcal{X}) \neq \mathcal{P}_{t+1}(\mathcal{X})$ and $\mathcal{P}_t(c|\mathcal{X}) = \mathcal{P}_{t+1}(c|\mathcal{X})$ are subcategories of concept drift, where covariate shift denotes changes in the distribution of the input data $\mathcal{X}$ only, excluding changes in the distribution of classes or labels (Moreno-Torres et al., 2012). Additionally, *real concept drift* refers to any changes in $\mathcal{P}(c|\mathcal{X})$, regardless of whether these changes are brought

on by variations in $\mathcal{P}(\mathcal{X})$. Label shift, concept shift, or conditional change are terms used to describe a subtype of real concept drift: $\mathcal{P}_t(\mathcal{X}) = \mathcal{P}_{t+1}(\mathcal{X})$ and $\mathcal{P}_t(c|\mathcal{X}) \neq \mathcal{P}_{t+1}(c|\mathcal{X})$. In this instance, the only distribution that varies over time is the distribution of the labels given $\mathcal{X}$, whereas the distribution of the input features alone does not. In simpler terms, *real concept drift* is the term used to describe scenarios with shifting boundaries (Gama et al., 2014).

To illustrate the difference between virtual and real concept drift, consider a recommendation system for sporting goods. The system's task is to classify sports products as relevant or irrelevant based on a buyer's interests. Initially, the buyer is interested in running shoes, making recommendations related to running gear appropriate, while bicycle equipment is considered irrelevant. Over time, new models of running shoes with updated features have become available. Although the specific attributes of the shoes have changed, they remain relevant to the buyer. This represents virtual drift, where the fundamental concept stays the same, but the characteristics of relevant items evolve.

However, if the buyer purchases a pair of running shoes and then decides to take up cycling instead of running, their interest shifts entirely from running gear to bicycle equipment. In this case, running shoes are no longer useful, and cycling gear becomes relevant. This shift in preference represents real concept drift, where the underlying concept itself changes.

In another example, think about a real estate-related online news feed. The user aims to divide the incoming news into relevant and irrelevant categories. Let's say the consumer is now looking for a new apartment. So, information about homes for habitation is pertinent, whereas news about vacation properties is not. If the news editor

changes, the wording will change, but the user will still find the dwellings relevant. This situation is equivalent to virtual drift. If, however, the editor, the writing style, and the user's preferences do not change, but because of a crisis, more articles about dwelling homes and fewer articles about vacation homes arise, this condition corresponds to a drift in the prior probabilities of the classes. On the other hand, if a user has already purchased a home and begins looking for a vacation spot, houses become less important and more important vacation spots. The scenario fits the real concept drift even while the writing style and prior probabilities remain constant.

Figure 2.1 also shows the connection between virtual and real concept drift. A data instance is represented by a dot in the illustration, and various class affiliations are indicated by various colours. Virtual drift occurs when the input data X changes but the decision boundary relative to the original data stays the same. Hence, the machine learning model does not need to be modified. In contrast, when there is real concept drift, the machine learning model's decision boundary needs to be modified to reflect the new class affiliations. While the distribution of the input data X remains unchanged in this particular instance, the concept drift illustrated indicates a label shift.

Finally, from a predictive standpoint, once a real concept drift happens, adaptation is necessary because the current decision boundary is out of date for the fresh arriving data (Gama et al., 2014). For the classification model to remain accurate, adaptation entails updating it for the new distribution.