Mixture-Based Clustering and Hidden Markov Models for Energy Management and Human Activity Recognition: Novel Approaches and Explainable Applications

Hussein Ghassan Ali Al-Bazzaz

A Thesis

in

Concordia Institute

for

Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Information and Systems Engineering) at

Concordia University

Montréal, Québec, Canada

April 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Hussein Ghassan Ali Al-Bazzaz**

Entitled: **Mixture-Based Clustering and Hidden Markov Models for Energy Management and Human Activity Recognition: Novel Approaches and Explainable Applications**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________ Internal Examiner and Chair
Dr. Chun Wang

______________________ External Examiner
Dr. Biao Li

______________________ Internal Examiner
Dr. Jamal Bentahar

______________________ Supervisor
Dr. Nizar Bouguila

______________________ Co-supervisor
Dr. Manar Amayri

Approved by ______________________
Zachary Patterson, Graduate Program Director
Concordia institute for Information Systems Engineering

__________ 2023

____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Mixture-Based Clustering and Hidden Markov Models for Energy Management and Human Activity Recognition: Novel Approaches and Explainable Applications

Hussein Ghassan Ali Al-Bazzaz, Ph.D.

Concordia University, 2023

In recent times, the rapid growth of data in various fields of life has created an immense need for powerful tools to extract useful information from data. This has motivated researchers to explore and devise new ideas and methods in the field of machine learning. Mixture models have gained substantial attention due to their ability to handle high-dimensional data efficiently and effectively. However, when adopting mixture models in such spaces, four crucial issues must be addressed, including the selection of probability density functions, estimation of mixture parameters, automatic determination of the number of components, identification of features that best discriminate the different components, and taking into account the temporal information. The primary objective of this thesis is to propose a unified model that addresses these interrelated problems. Moreover, this thesis proposes a novel approach that incorporates explainability.

This thesis presents innovative mixture-based modelling approaches tailored for diverse applications, such as household energy consumption characterization, energy demand management, fault detection and diagnosis and human activity recognition. The primary contributions of this thesis encompass the following aspects:

Initially, we propose an unsupervised feature selection approach embedded within a finite bounded asymmetric generalized Gaussian mixture model. This model is adept at handling synthetic and real-life smart meter data, utilizing three distinct feature extraction methods. By employing the expectation-maximization algorithm in conjunction with the minimum message length criterion, we are able to concurrently estimate the model parameters, perform model selection, and execute feature selection. This unified optimization process facilitates the identification of household electricity consumption profiles along with the optimal subset of attributes defining each profile. Furthermore, we investigate the impact of household characteristics on electricity usage patterns to pinpoint households that are ideal candidates for demand reduction initiatives.

Subsequently, we introduce a semi-supervised learning approach for the mixture of mixtures of bounded asymmetric generalized Gaussian and uniform distributions. The integration of the uniform distribution within the inner mixture bolsters the model's resilience to outliers. In the unsupervised learning approach, the minimum message length criterion is utilized to ascertain the optimal number of mixture components. The proposed models are validated through a range of applications, including chiller fault detection and diagnosis, occupancy estimation, and energy consumption characterization. Additionally, we incorporate explainability into our models and establish a moderate trade-off between prediction accuracy and interpretability.

Finally, we devise four novel models for human activity recognition (HAR): bounded asymmetric generalized Gaussian mixture-based hidden Markov model with feature selection (BAGGM-FSHMM), bounded asymmetric generalized Gaussian mixture-based hidden Markov model (BAGGM-HMM), asymmetric generalized Gaussian mixture-based hidden Markov model with feature selection (AGGM-FSHMM), and asymmetric generalized Gaussian mixture-based hidden Markov model (AGGM-HMM). We develop an innovative method for simultaneous estimation of feature saliencies and model parameters in BAGGM-FSHMM and AGGM-FSHMM while integrating the bounded support asymmetric generalized Gaussian distribution (BAGGD), the asymmetric generalized Gaussian distribution (AGGD) in the BAGGM-HMM and AGGM-HMM respectively. The aforementioned proposed models are validated using video-based and sensor-based HAR applications, showcasing their superiority over several mixture-based hidden Markov models (HMMs) across various performance metrics. We demonstrate that the independent incorporation of feature selection and bounded support distribution in a HAR system yields benefits; Simultaneously, combining both concepts results in the most effective model among the proposed models.

Acknowledgments

I would like to first thank Professor Nizar Bouguila for his continuous help and support. He has helped me develop a lot as a researcher through his continuous provision of effective research resources. Professor Bouguila will always be my idol while pursuing a career in academia regarding research ethics and innovation.

I want to thank Professor Manar Amayri for her valuable guidance, comments, and provision of research resources.

Many thanks to the rest of my Ph.D. committee members, namely Dr. Chun Wang, Dr. Biao Li, and Dr. Jamal Bentahar, for their guidance and comments, which have significantly helped to improve my research papers.

I would also like to thank my mentor and my best friend, with whom I have spent many years learning from their research expertise, Dr. Muhammad Azam.

I would not forget to thank my friend, my teacher, and the engineer Muhammad Al-Sheikhly for supporting me and guiding me throughout my studies in Lebanon and Canada.

Naturally, the best of my thanks go to my Mother (Athraa Zaki Jassim Al-Khshali) and my Father (Ghassan Ali Abdulkareem Al-Bazzaz) for their immeasurable sacrifices that gave me the opportunity to be here and contribute as a researcher in my passion, Artificial Intelligence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the past decade, there has been a remarkable growth in the amount and dimensionality of data due to technological advancements. It is crucial to find ways to model and analyze such multidimensional data. The process of modelling and extracting useful insights from multidimensional data relies on the ability to identify complex patterns, regularities, and relationships within the data. In recent years, several algorithms have been developed with the aim of automatically learning to recognize these complex patterns and making intelligent decisions based on observed data. Machine learning, a branch of artificial intelligence, provides a systematic approach to developing and studying automatic techniques capable of learning models and their parameters from training data [2–5]. Machine learning and statistical pattern recognition have witnessed remarkable development in recent years due to their extensive applications in various fields, including engineering, medicine, computer science, psychology, neuroscience, physics, and mathematics [6, 7]. The emergence of novel supervised [8–10] and unsupervised methods [11], as a result of recent advances in machine learning, has attracted the attention of researchers from different domains, owing to their potential to support the modelling and analysis of different data.

The unsupervised partitioning of data is considered in a wide range of tasks that involve grouping observations into clusters or parts that are similar in composition and different from each other. The categorization of patterns and objects based on similarities is a crucial aspect of most real-life applications and is commonly known as clustering or cluster analysis. There are various approaches to clustering, including hierarchical, relocation, probabilistic, density-based, and grid-based methods, which can be employed according to the real-life application.

Hierarchical methods generate clusters incrementally, leveraging the connectivity matrix that represents the similarity among data items. There are two primary hierarchical clustering strategies: agglomerative and divisive. The agglomerative approach commences with singleton clusters, each containing a single element, and progressively merges cluster pairs. Conversely, the divisive method starts with a comprehensive cluster containing all objects and iteratively divides it into distinct clusters.

Relocation algorithms, unlike hierarchical techniques, do not develop clusters step by step. Instead, they initiate with a random partition and iteratively reposition data items among existing clusters to optimize them. Typically, these methods necessitate a predetermined number of clusters. The most prevalent relocation algorithm is the K-means approach, which alternates between two iterative steps: data assignment and centroid value updating.

Probabilistic clustering techniques are based on the assumption that the dataset represents a sample independently drawn from multiple populations. Density-based clustering identifies clusters as high-density regions in the feature space, separated by low-density areas. This approach excels in detecting clusters of various shapes, as it utilizes the concepts of density and connectivity to account for local data distribution and requires defining neighbourhood data and nearest neighbour computations.

Grid-based clustering algorithms partition the feature space, aggregating dense neighbouring segments. A segment is a multi-rectangular region in the feature space, resulting from the Cartesian product of individual feature sub-ranges. Consequently, data partitioning is achieved through space partitioning. Some grid-based techniques prune the attribute space a priori, performing subspace clustering, which is crucial for high-dimensional data when irrelevant features may obscure the clustering tendency. This method can be regarded as an extension of traditional clustering that aims to identify clusters in various sub-spaces within a dataset.

In this thesis, I will primarily focus on probabilistic approaches, specifically mixture models. Mixture models, as a machine learning technique, have garnered significant interest across various applications. They are typically employed to model intricate datasets by postulating that each observation originates from one of several distinct groups or components [7]. Furthermore, mixture models have demonstrated success in tasks like clustering and density estimation [12, 13].

1.1 Mixture Models: An Introduction

Mixture models consist of convex combinations of two or more probability density functions (PDFs). By amalgamating the attributes of the individual PDFs, mixture models possess the ability to approximate any arbitrary distribution [14]. As a result, selecting the most appropriate PDF to accurately represent the mixture components is of paramount importance, as it directly impacts the model's capability to characterize the shape of the data [15, 16]. Furthermore, an ill-suited choice of PDF may necessitate the mixture model to increase the number of components in order to fit the data accurately, which could lead to overfitting. The mixture of Gaussians is among the most fundamental and widely used statistical models, frequently justified for asymptotic reasons, meaning that the sample is assumed to be adequately large [17]. However, it has been observed that Gaussian distributions are generally unsuitable for modelling data in intricate real-life applications [18]. For example, natural image clutter tends to be non-Gaussian in nature.

The Gaussian mixture model is a popular choice as an isotropic probability density function that can compactly model and represent the intrinsic grouping of the data. It comprises a set of parameters, including a mean vector and a covariance matrix, that describe each discovered pattern's properties. The EM algorithm has been widely used to estimate the Gaussian mixture model's parameters that best fit the data in many applications [19, 20]. However, the Gaussian distribution's limitations have been observed in various real-life applications. For instance, it has a rigid bell shape and a short tail, which is unsuitable for most real-world problems [21]. Additionally, the Gaussian distribution is unbounded with a support range that extends from $-\infty$ to ∞ and is symmetric around its mean [22, 23]. The data clusters of real-life applications are usually bounded and most likely have a non-Gaussian density, signifying a density function that is asymmetric, non-bell shaped, or both [24–27]. The Gaussian distribution's fixed kurtosis makes the mixture model vulnerable to individual data cluster outliers, and the distribution is unsuitable for assigning a relatively low probability of occurrence to the individual class outliers [28, 29]. Therefore, many applications that incorporate the Gaussian mixture model seek outlier detection and removal techniques within their workflow, which incurs additional computational expenses [30–33].

Several studies have demonstrated that the generalized Gaussian distribution (GGD), can be an effective alternative to Gaussian distributions due to its shape adaptability [16, 34]. This flexibility allows the modelling of a broad array of non-Gaussian signals [35–38]. The GGD encompasses the Laplacian, Gaussian, and asymptotically uniform distributions as special cases [39]. It has been employed in various challenging problems, demonstrating its versatility and effectiveness in diverse contexts (e.g., [40–42]). However, the GGD is still a symmetrical distribution, rendering it unsuitable for modelling non-symmetrical data. In light of this limitation, we propose examining two non-symmetrical distributions throughout the remainder of this thesis: the asymmetric generalized Gaussian and its bounded variant. By considering these distributions, we aim to enhance the modelling capabilities of mixture models, allowing for a more accurate representation of complex data with non-symmetrical characteristics. This exploration will contribute to a deeper understanding of mixture models' strengths and limitations, fostering further advancements in the field.

The asymmetric Gaussian distribution (AGD) builds upon the foundation of the Gaussian distribution, offering a more nuanced representation of data by addressing the asymmetry often present in real-life datasets. While the Gaussian distribution assumes symmetric data around the mean, the asymmetric Gaussian allows for different standard deviations on each side of the mean, capturing the skewed nature of many empirical observations. Despite this advantage, the asymmetric Gaussian distribution still maintains the fundamental bell-shaped curve of the Gaussian distribution. By accommodating asymmetry while preserving the Gaussian structure, the asymmetric Gaussian distribution serves as a valuable tool for handling a variety of data analysis tasks that involve skewed data patterns.

The asymmetric generalized Gaussian distribution (AGGD) [15] presents a significant improvement over the generalized Gaussian distribution (GGD) and the AGD in various aspects. While the GGD is known for its shape flexibility, it remains a symmetrical distribution, which limits its ability to accurately model non-symmetrical data. In contrast, the AGGD addresses this limitation by incorporating asymmetry in the distribution like AGD, while allowing for a more precise representation of complex data with non-symmetrical characteristics like GGD. This enhancement enables the AGGD to capture a broader range of distribution shapes, which proves beneficial in diverse applications where data exhibit varying degrees of skewness. Moreover, the AGGD maintains the advantages of the GGD, as it can generalize to the GGD when the left standard deviation is equal to the right standard deviation. This dual capacity of modelling both symmetrical and asymmetrical data makes the AGGD a more versatile and powerful tool in the realm of mixture models.

Bounded probability density functions offer an essential extension to the traditional probability density functions as discussed in [43], such as the generalized Gaussian distribution (GGD) and asymmetric generalized Gaussian distribution (AGGD). These bounded variants are particularly useful for modelling real-world data that often exhibit constraints or limits within a specific range as demonstrated in [21], [44], and [22, 44]. By incorporating boundaries into the distribution, these functions can more accurately represent real-world data that naturally fall within a finite interval. Examples of such data include percentages, proportions, and measurements with known upper and lower bounds. Bounded versions of the GGD and AGGD provide additional flexibility by capturing the characteristics of both symmetrical and asymmetrical data while adhering to the imposed constraints. Consequently, they prove to be invaluable tools in various applications, as they offer a more precise representation of data distributions and improve the overall performance of mixture models in capturing the underlying structure of the data.

A mixture of mixture models is a statistical modelling technique used to describe complex data with heterogeneous sub-populations. It involves combining multiple mixture models, where each component of a mixture model is itself a mixture model. The inner mixture models are used to capture finer details of the sub-populations within each component, while the outer mixture models are used to model the overall distribution of the data. The incorporation of inner mixtures within each mixture model's component allows for greater flexibility in modelling the data, as it enables the identification and characterization of finer-grained sub-populations within each component. This is particularly useful when dealing with complex data that cannot be adequately described by a single mixture model. By incorporating inner mixtures, the mixture of mixture models can provide a more accurate and detailed representation of the data, making it a powerful tool for a wide range of statistical applications, such as clustering, classification, and density estimation.

In this thesis, we will explore the probabilistic aspects of mixture models in greater depth, investigating their strengths and limitations as well as their applications across various domains. By examining the mathematical foundations and practical implementations of these models, we aim to provide a comprehensive understanding of their role in contemporary machine learning research. This deeper understanding will contribute to the ongoing development and refinement of mixture models, enabling further advancements in the field.

1.2 Parameter Learning of Mixture Models

Unsupervised learning offers the potential to discover hidden patterns and structures within complex data without relying on labelled examples, making it a more scalable and versatile approach for various real-world applications. In contrast, supervised learning, especially when utilizing neural networks [45], often suffers from high dependence on large labelled datasets and is prone to overfitting, which limits the applicability and generalization capabilities of these models in dynamic and less-structured environments. Mixture models are widely used to represent complex data distributions by combining multiple probability density functions (PDFs). The unsupervised learning of mixture models is crucial to ensure the effectiveness of the model in various applications. There are several methods to fit mixture models to training data, such as maximum likelihood estimation [46], Bayesian variational inference [47–50], and Bayesian MCMC inference [51–54]. In this section, we discuss the key techniques employed for parameter learning in mixture models, emphasizing the maximum likelihood estimation (MLE) using the Expectation-Maximization (EM) algorithm, the Minimum Message Length (MML) criterion, and the semi-supervised learning approach.

The EM algorithm is a popular iterative method for MLE in mixture models. It is particularly effective in handling incomplete or missing data and has been widely adopted for parameter estimation in Gaussian mixture models and other complex distributions. The EM algorithm consists of two main steps: the Expectation (E) step and the Maximization (M) step. In the E-step, the algorithm calculates the expected values of the latent variables given the observed data and current parameter estimates. In the M-step, it updates the parameter estimates by maximizing the expected complete-data log-likelihood obtained in the E-step. The algorithm iterates between these two steps until convergence, ultimately yielding the maximum likelihood estimates of the mixture model parameters.

The MML criterion is an information-theoretic model selection technique that aims to minimize the message length needed to encode both the model and the data. By incorporating the MML criterion within the EM algorithm, it is possible to simultaneously achieve feature saliency and determine the optimal number of mixture components. In this approach, the optimization process not only estimates the parameters of the mixture model but also identifies the actual number of components and the optimal subset of attributes that define each component. This integrated optimization strategy allows for more efficient and effective parameter learning in mixture models, ensuring that the resulting model accurately represents the underlying data structure.

In many real-world applications, obtaining fully labelled data is challenging and costly, making the use of semi-supervised learning techniques highly desirable. The semi-supervised approach to maximum likelihood estimation for mixture models leverages both labelled and unlabeled data to improve the accuracy and robustness of parameter estimation. By incorporating the available labelled data, the semi-supervised MLE can guide the learning process and refine the model parameters more effectively than unsupervised approaches. This results in a better representation of the underlying data distribution and ultimately improves the performance of the mixture model in various applications.

Within this thesis, Chapters 2 and 3 showcase the implementation of unsupervised and semi-supervised maximum likelihood estimation through the EM algorithm for several mixture models.

1.3 Feature Selection

In the realm of statistical modelling, and specifically finite mixtures, feature selection is a critical aspect, as it involves identifying relevant or discriminative features that describe the data, particularly in the analysis of high-dimensional data that has been extensively researched in the past. The primary objective is not only to determine mixture components and their parameters but also to offer the most parsimonious model that can accurately depict the data. It is important to note that humans' approach to clustering and recognition involves selecting a few key features (i.e., only relevant information is considered while ignoring irrelevant information [55, 56]) and then clustering the data based on these features [57]. Additionally, feature selection can accelerate learning and enhance model accuracy and generalization. As a result, feature selection has proven to be a vital step in various applications such as image processing, computer vision, and pattern recognition, including object detection [58], handwriting separation [59], and image retrieval, categorization, and recognition [60].

Nonetheless, most mixture model research presumes that all features have equal weight and employs a pre-processing step like principal components analysis (PCA) to convert the original features into a reduced-dimension space. This approach's primary drawback is the loss of the original features' physical meaning [61]. Moreover, the quality of the features used significantly impacts the learning of mixture parameters (i.e., both model selection and parameter estimation), as demonstrated, for example, in [62], leading to a renewed focus on the feature selection problem, particularly in unsupervised settings. Feature selection was achieved in several prior publications simultaneously to the parameter estimation in maximum likelihood estimation [15, 63] and even Bayesian approaches [64, 65]. Like many other model-based feature selection approaches (e.g., [66]), this work is based on the Gaussian assumption, assuming diagonal covariance matrices [66] for all clusters (i.e., all features are considered independent). In this thesis, following recent approaches (e.g., [60, 62]), we attempt to address the feature selection problem in unsupervised learning by framing it as an estimation problem, thus avoiding any combinatorial search. We assign a relevance weight to each feature, which measures its dependence on class labels.

1.4 Explainable Artificial Intelligence

Recent advancements in machine learning algorithms have enabled the detection of patterns in data that were previously impossible to identify. However, the lack of transparency in the decision-making processes of these algorithms has raised concerns about their reliability and trustworthiness. This has led to an increasing interest in the field of explainable artificial intelligence (XAI) [67–69].

While there have been significant research efforts to enhance the interpretability of data-driven models, most prior works have focused on post-modelling explainability, which has several limitations [70]. Firstly, it does not provide any insights into the training data used to build the model. Secondly, it heavily relies on the specific model being used.

In recent years, the term ”explainability” has gained widespread attention, and there has been an exponential increase in its usage in research publications [71, 72]. Our proposed model aims to address the limitations of prior works by providing pre-modelling explainability in the context of model-based classification and clustering. Specifically, we develop an approach that not only recognizes different patterns within the data and the statistical properties of each pattern, but also defines the boundaries between the discovered patterns in terms of important data attributes [**?**, 70].

Our proposed method integrates a mixture model with a decision tree (DT) algorithm to provide explainable predictions that generalize well to unseen data. The DT algorithm is used to train a small binary threshold tree, with the number of leaves equal to the number of clusters assumed in the mixture model. The predictions of the DT algorithm are easily interpretable using simple If-Then rules. By integrating the DT algorithm with our proposed mixture model, we can provide insights into the model and its prediction by defining the boundaries between the detected patterns in terms of important attributes [72].

We demonstrate the integrated explainability within our proposed framework by applying it to chiller fault detection and diagnosis, energy consumers’ categorization, and occupancy estimation. In the case of chiller fault detection, we aim not only to classify faults but also to provide interpretability in the form of If-Then statements using the values of the data features. Similarly, in the case of energy consumers’ categorization, we aim to provide interpretability in the form of simple If-Then statements with specific values of high-level features to define the boundaries between energy consumers. Our proposed model can help utility companies to identify suitable households for demand reduction initiatives such as demand response and energy efficiency, thereby reducing costs and greenhouse gas emissions [73–77].

As will be demonstrated in Chapter 3, the integration of mixture models with explainable artificial intelligence offers a powerful combination that addresses some of the key challenges faced in machine learning. Mixture models are generative models that excel in generalizing to unseen data based on a sample population, while explainable AI focuses on making the inner workings of complex models transparent and understandable. Decision trees, as supervised learning algorithms, are known to be prone to overfitting. By adding explainability, we can create an interpretable model that not only generalizes well to unseen data but also delivers comprehensible results through straightforward if-then statements. This fusion of approaches enables more accurate predictions, and fosters trust and transparency in the underlying machine-learning models.

1.5 Hidden Markov Models

Over the past ten years, Hidden Markov Models (HMM) have garnered significant attention from researchers due to their expanded capabilities beyond the originally investigated speech-related tasks [78]. Applications of HMMs now include handwritten character recognition, music analysis, stock market prediction, earthquake forecasting, video categorization, security monitoring, and network evaluation. HMMs are probabilistic models that belong to the generative category of machine learning algorithms. In machine learning, data modelling techniques typically fall into one of two main categories: discriminative or generative. Generally, discriminative models are trained to determine a connection between input data and class labels, while generative models first learn the distribution of the classes before making predictions. Mathematically, discriminative models represent the probability of a class label given the input data, while generative models denote the joint probability of both input data and class labels, which is used to calculate the appropriate probability for classification. In a way, HMMs can be viewed as an extension of mixture models that incorporates the temporal dimension. This means they have the ability to perform spatiotemporal modelling, taking into account both spatial and temporal features. Therefore, the HMMs are among the most widely used statistical techniques for probabilistic modelling of sequential and time series data [79, 80]. An HMM is a highly-regarded dual stochastic model that leverages a concise set of features to extract underlying statistics [78]. The structure primarily consists of a Markov chain of hidden variables, each linked to a conditional observation. A Markov chain offers one of the simplest ways to represent sequential patterns in time series data and was first introduced by Andrey Markov in the early 1900s. The late 1960s and early 1970s witnessed a surge in publications by Leonard E. Baum and other researchers that explored and addressed its statistical methods and modelling [79]. This approach allows us to maintain generality while easing the assumption of independent and identically distributed variables [81].

In mathematical terms, an HMM is defined by an underlying stochastic process consisting of a number of hidden states that form a Markov chain. Each state is regulated by an initial probability, and the transitions between states at a given time t can be represented by a transition matrix. Within each state s_t, at time t an observation is produced according to its distribution, which may be either discrete or continuous. This forms the set of observable stochastic processes. Conversely, the specific parameters of a probability distribution determine the observation emission for a continuous observed symbol sequence. The Gaussian distribution is most frequently employed as the emission distribution [79, 82, 83].

The mixture-based Hidden Markov Model (MM-HMM) is an advanced variation of the traditional HMM that offers increased modelling capabilities for complex data. This innovative approach incorporates a mixture of probability distributions within each hidden state, providing a more flexible and accurate representation of the underlying data-generating process. By integrating multiple distributions, MM-HMMs can better capture intricate patterns and relationships within time series or sequential data. As a result, mixture-based HMMs have emerged as a powerful tool for various applications, including speech recognition, financial forecasting, bio-informatics [84], human action recognition [85], and many others that require robust modelling of complex, dynamic phenomena. Human action recognition, in particular, benefits from the enhanced ability of MM-HMMs to discern subtle variations in movement patterns, making them a valuable asset in areas such as video surveillance, human-computer interaction, and sports analytics.

The Baum-Welch algorithm for Mixture-based Hidden Markov Models (MM-HMMs) is an iterative optimization technique employed to estimate the model parameters. Also known as the forward-backward algorithm, it leverages the EM framework to refine the parameters of the underlying mixture distributions within each hidden state. The algorithm consists of two primary components: the forward algorithm and the backward algorithm. These methods are executed recursively, forming the complete algorithm that ensures convergence towards more compact clusters at each iteration [86]. Starting with an initial random data clustering, the Baum-Welch algorithm for MM-HMMs guarantees convergence and aims to optimize the model in order to better capture the complexities of the sequential data under analysis. The process terminates when there are no significant changes in the log-likelihood ratios, ultimately yielding an optimized MM-HMM [86].

The Viterbi Algorithm is a dynamic programming-based method that is integral to HMMs and other associated probabilistic models [87]. Introduced by Andrew Viterbi in 1967 [88], its primary purpose is to identify the most probable sequence of hidden states, known as the Viterbi path, given a sequence of observations. The algorithm efficiently computes the optimal state sequence by utilizing a recursive process that maximizes the joint probability of both the observations and the corresponding hidden states.

In the context of a specific HMM, the Viterbi algorithm aims to determine the most likely progression of states that generated a given observation sequence, thereby addressing the decoding problem. This involves selecting the most probable states at each individual time step t, which in turn maximizes the expected number of correct separate states. The Viterbi Algorithm has been widely applied across various fields, including speech recognition, natural language processing, bio-informatics, and other areas that necessitate decoding and inference in the presence of hidden information.

In Chapter 4, we introduce four new models for human activity recognition, namely BAGGM-FSHMM, BAGGM-HMM, AGGM-FSHMM, and AGGM-HMM. Each of these models represents a unique combination of techniques designed to enhance the performance of Hidden Markov Models in this application domain. BAGGM-FSHMM stands for Bounded Asymmetric Generalized Gaussian Mixture-based Hidden Markov Model with Feature Selection. Similarly, BAGGM-HMM represents the Bounded Asymmetric Generalized Gaussian Mixture-based Hidden Markov Model, without feature selection. AGGM-FSHMM denotes the Asymmetric Generalized Gaussian Mixture-based Hidden Markov Model with Feature Selection, while AGGM-HMM refers to the Asymmetric Generalized Gaussian Mixture-based Hidden Markov Model without feature selection. These models aim to improve the accuracy and efficiency of human activity recognition by incorporating various advanced techniques and methodologies.

1.6 Contributions

The aim of this thesis is to propose several novel models for recognizing patterns within data with and without taking into account the temporal axis. The models proposed utilized effective frameworks that model the different ways random variables change for a given pattern.

- ☞ **The bounded asymmetric generalized Gaussian mixture model with minimum message length criterion:** The bounded asymmetric generalized Gaussian mixture model was extended within a framework that estimates model parameters and feature saliencies simultaneously. The proposed model has been proven effective in modelling smart meter data within different real and synthetic datasets. This novel framework was submitted as a journal research manuscript with the title ”A mixture-based clustering approach for household energy consumption segmentation and feature weighting” to the Journal of Sustainable Energy, grids, and Networks.

- ☞ In order to incorporate further robustness within mixture models. We incorporate a Uniform distribution within an inner mixture of a mixture of mixtures that further extend the bounded asymmetric generalized Gaussian mixture model. A semi-supervised learning setting is also proposed for the novel mixture of mixtures to reduce the reliance on labelled data in fault detection and diagnosis and occupancy estimation applications using real-world datasets. These novel frameworks have been published as research papers as follows:

  - A journal research manuscript was submitted with the title ”Explainable finite mixture of mixtures of bounded asymmetric generalized Gaussian and Uniform distributions learning for energy demand management” to the following publication ”ACM transactions on intelligent systems and technology”.
  - A research paper was submitted as a research manuscript with the title ”Explainable robust Smart Meter Data Clustering for Improved Energy Management” to the following conference ”IEEE International Conference on Systems, Man, and Cybernetics (SMC)”

- ☞ Four mixture-based hidden Markov models were proposed for a human activity recognition application. Using these novel models, the following attempts at publications were made:

- A journal research paper was submitted with the title ”Advanced Models for Human Activity Recognition using Mixture-Based Hidden Markov Models with Feature Saliencies” to the following publication, ”Engineering Applications of Artificial Intelligence”.
- A research manuscript with the title ”Enhancing Human Action Recognition with Asymmetric Generalized Gaussian Mixture Model-Based Hidden Markov Models and Bounded Support” was submitted to the following conference ”IEEE International Conference on Systems, Man, and Cybernetics (SMC)”

1.7 Thesis Overview

This thesis is organized as follows:

❑ Chapter 1: This chapter introduces basic concepts that were used in this thesis to model real-world data, such as mixture models, hidden Markov models and their optimization approaches.

❑ Chapter 2: This study proposes a learning framework that involves the expectation-maximization algorithm (EM) and the minimum message length (MML) criterion for a simultaneous feature and model selection approach in the context of the bounded asymmetric generalized Gaussian mixture model. The proposed algorithm demonstrates superior clustering efficacy compared to several state-of-the-art clustering algorithms in the analysis of high-resolution smart meter data.

❑ Chapter 3: This study introduces a mixture of mixtures of bounded asymmetric generalized Gaussian and uniform distributions and proposes model-based classification and clustering algorithms. The proposed algorithm's predictions are interpretable to the user's perspective through simple If-Then statements using a small binary decision tree, increasing the credibility of the algorithm's predictions. The proposed algorithm demonstrates its reliability and superiority to several state-of-the-art machine learning algorithms in real-world applications.

❑ Chapter 4: This study proposes an asymmetric generalized Gaussian mixture-based hidden

Markov model (AGGM-HMM) and a bounded asymmetric generalized Gaussian mixture-based hidden Markov model (BAGGM-HMM) for recognizing human actions from a sequence of observations. The proposed models leverage the flexible and robust asymmetric generalized Gaussian distribution and the bounded variant to model real-life applications' data. The proposed frameworks demonstrate superiority to several state-of-the-art human activity recognition methods in sensor-based and video-based recognition scenarios.

- Conclusion: In this chapter, we provide a summary of our contributions. Additionally, we discuss potential avenues for future research.

Chapter 2

Mixture-Based Clustering Approach For Household Energy Consumption Segmentation and Feature Weighting

In the previous chapter, we introduced several important topics in machine learning, including mixture models in Section 1.1 and their learning criteria in Section 1.2. In this chapter, important concepts shall be used to introduce a novel mixture model and its learning criteria as a solution for an important real-life problem. Recently, the intervals of publicly available smart meter data have become as small as one second compared to one month in earlier times. Understanding these variations supplies the opportunity to discover important information for several capabilities involving metering data. As a data mining method, clustering analysis has been widely used to discover unique energy consumption patterns and trends and the consumers that follow them. The high-resolution smart meter data present several challenges that recent studies have failed to address, such as non-Gaussian data shape, an unknown number of clusters, and high dimensional feature space with variable importance. Although several studies have addressed the problems of an unknown optimal number of clusters and optimal feature subset in modelling smart meter data independently, these problems are interrelated and must be addressed simultaneously. Therefore, this chapter proposes a learning framework that involves the expectation-maximization algorithm (EM) and the minimum message length (MML) criterion for a simultaneous feature and model selection approach in the context of the bounded asymmetric generalized Gaussian mixture model. Our experiments attempt to simulate an efficient analysis scenario of smart metering data with three feature extraction methods. The performance of the proposed algorithm is compared against the performance of several state-of-the-art clustering algorithms. We validate the clustering efficacy of the proposed algorithm with several performance measures using two synthetic and real smart meter datasets. The resulting clusters characterize the variations in residential energy consumption to help accurately determine the group of suitable households for utility companies' demand reduction initiatives.

2.1 Introduction

The adaptation of Advanced Metering Infrastructure (AMI) in Europe has been a significant contributor to the overachievement of the energy efficiency gains of the EU 20-20-20 energy policy. Following the success in Europe, smart meters have been deployed globally in countries attempting to modernize their electricity networks. As a result, these advancements in energy metering technologies have generated a wealth of new electrical power consumption records with an adequate and consistent frequency within the residential sector. Several smart meter datasets were made publicly available in [89–93] with a granularity of 1-min to 1-h. These datasets dramatically increase the amount of electricity use information over once-per-month meter reads. The high-resolution smart meter data analysis reveals several insights that were not previously possible. Thus, they provide a unique opportunity to understand a household's energy consumption pattern. The information obtained from such patterns can potentially enhance the targeting and customization of metering data capabilities such as demand response (DR), energy efficiency (EE), load forecasting and pricing intelligence programs and the improvement of energy-saving recommendations [94–96]. DR is an incentive program that enables the possibility for utility companies to save money on unnecessary investments and lower emissions of greenhouse gases (GHG). DR induces households to reduce their energy consumption levels at times of high wholesale market prices or when system reliability is jeopardized. EE programs aim to reduce the power demand of households while maintaining their consumption habits.

The classical classification of energy consumers that is based on the explicit type of activity does not correlate well with the evolution of energy consumption [97], rendering the efforts to increase energy efficiency ineffective. Therefore, identifying electricity customers with similar energy consumption patterns is perhaps significantly more helpful [98, 99]. In order to transform smart meter records into valuable information taking part in customer groupings, traditional machine learning exploratory analysis tools such as unsupervised learning techniques are utilized [94]. Clustering is a statistical data analysis technique that can uncover or infer intrinsic properties and group the data into several components according to the observations' similarities. As a soft clustering approach, the Gaussian mixture's reliability and minimum impact on computational capabilities have made it a good candidate for modelling smart meter data. However, it has the following weaknesses: the distribution has a fixed shape; preventing it from generalizing to different classes of distributions [100], the distribution's tails are short; making it vulnerable to outliers and therefore is not the best choice for mixture models in real applications [101]. The Gaussian distribution does not fit data well within a mixture model if the data has an asymmetric distribution, as demonstrated in Figure 2.1. Additionally, estimating the data's bounded support region in mixture models has achieved improved performance in diverse real-life applications [21, 25, 26, 43, 102, 103].

The deployment of AMI has introduced high dimensionality in modern energy consumption datasets. As an example, in the Irish smart meter trial [89], with a reading interval of 30 minutes, an energy consumer's load curve consists of 25728 features in the raw format of the dataset. Patterns are easily distinguished within observations that are represented with features of high entropy. However, in practice, the best clustering performance is achieved with the best set of features that concisely describe the load curve. Feature selection has several advantages: it is well established to improve the performance of model-based categorization [104], and it helps develop interpretable models that are reduced in complexity within applications across several disciplines. The search for the optimal number of clusters and the optimal set of features is an interrelated optimization problem. However, searching for the optimal set of features is challenging in an unsupervised setting because there is no clear criterion for the optimization process since the number of clusters is unknown. Historically, in order to find the optimal number of features, an exhaustive search is done through the space of all feature subsets [105–107]. Additionally, non-exhaustive search techniques do not guarantee finding the optimal feature subset. Therefore, an efficient solution was proposed within an unsupervised setting [62]; the optimal feature subset search is converted into an estimation problem parallel to the learning of mixture models where a vector of feature weights is estimated using the EM algorithm.

In this proposal, we develop a finite bounded asymmetric generalized Gaussian mixture model with simultaneous feature selection and model selection, generalizing to an extensive range of mixture models to model smart meter data in an unsupervised manner. The proposed algorithm finds the model's optimal subset of features, optimal count of components, and optimal parameters using the EM algorithm and the MML model selection criterion. Smart meter data classes consist of observations of continuous random variables, which makes it ideal for modelling using the families of continuous probability distributions. Additionally, the data is ideal to be used to demonstrate the generalization of the proposed model due to the various feature extraction methods used in prior works to represent this data. As the representation of the data changes, so does its distribution's skewness, bounds and shape. This chapter will demonstrate how our proposed FSBAGGMM generalizes to a wide range of mixture models, including the bounded variants. The generalization and the superiority of our mixture model are evident in the experimental analysis. In our experimental analysis, our proposed method outperforms the following: asymmetric generalized Gaussian mixture model-based feature selection (FSAGGMM), Bounded Asymmetric Generalized Gaussian Mixture Model (BAGGMM), and the Asymmetric Generalized Gaussian Mixture Model (AGGMM) according to several performance evaluation metrics. Additionally, our proposed mixture model has been implemented using Concordia University's High-Performance Computing (HPC) Facility: Speed [108].

2.1.1 Prior Work

Several applications were proposed to make use of energy consumption records. Additionally, due to the utilization of smart meters, the feasibility and reliability of such applications have increased. *Non-Intrusive Load Monitoring* (NILM), for example, has made heating, ventilation, and air conditioning (HVAC) fault detection applications more effective and reliable. The behaviour of the HVAC system is identified and monitored, relying on smart meter readings instead of installed

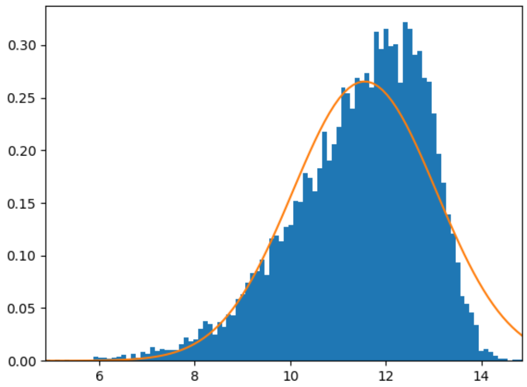


Figure 2.1: The Gaussian distribution Symmetry Problem

sensors [109]. An example of a system that uses smart meter records as input and produces load forecasting and recommendations for better energy efficiency is presented in [110]. A smart meter analytics solution was proposed to present customers with a user-friendly web portal they can use to understand their bills [111]. Energy consumption records were also used in applications to conclude consumption predictions and recommendations [73, 110, 112].

Clustering has proven helpful to group low, and high-voltage customers [113, 114]. Additionally, demand management programs have successfully utilized clustering in order to select suitable candidate energy consumers [75–77]. Thus, several approaches have been employed for the segmentation of energy users, such as K-Means [112], Euclidean distance-based clustering [114], and multi-resolution clustering in spectral-domain [115]. Similarly, several clustering methods such as hierarchical clustering, K-means, fuzzy K-means, and Self-organizing maps (SOM) have been used to group consumers with similar energy consumption patterns in [113]. SOM was tested for its capability to classify consumption profiles in [116]. Clustering has also proven useful to enhance energy consumption prediction using a two-layer feed-forward artificial neural network [96]. The Gaussian mixture model, optimized by the EM algorithm, was utilized in [117] and [73] as a non-distance-based consumer segmentation tool. Other finite mixture models have also been used within the context of the same application [118, 119].

In order to model smart meter data in different representations, several limitations imposed by the Gaussian mixture model must be overcome. Several distributions have been used as a base distribution of mixture models to overcome the shape rigidity of the Gaussian distribution, such as the Student's-t distribution [120–122] and the *generalized Gaussian distribution* (GGD) [38, 40, 123]. Compared to the Gaussian distribution, the Student's t distribution has an additional parameter (ν) called the degree of freedom that allows the distribution to generalize to different probability distributions. The Student's t distribution is identical to the Cauchy distribution when ($\nu = 1$) and approaches the Gaussian distribution as (ν) approaches infinity. As for the GGD, the additional parameter per component (λ) is called the shape parameter; it controls the tails of the distribution, making it far more flexible to different types of data and less vulnerable to outliers [36, 37, 124]. In more recent studies, the AGGD was used as a base distribution for mixture models [63, 125]. The AGGD can generalize to a large class of distributions such as the Impulsive, the Laplacian, the Gaussian, and the uniform distribution, in addition to the ability to fit asymmetric data [126]. Additionally, and in order for mixture components to fit better to real-life data, the bounded support concept was adopted in several finite mixture models [24–26].

Several feature extraction methods were utilized to process high dimensional data in electrical load observations and turn it into a new set of reduced feature space. In [127], a scalable algorithm for data processing has been proposed for a dataset collected from 10,000 Australian homes over a year. Dimensionality reduction is accomplished by employing a sparse representation technique in [128]. An encoding system has given representations for energy consumers with a pre-processed dictionary in [74]. The discovery of prominent energy consumption time windows is crucial for feature extraction and, therefore, modelling the typical consumer's behaviour. Through a thorough analysis of several smart meter trials, researchers have been able to identify four time periods where the most extensive distribution of peak demands occurs within smart meter datasets [73]. The energy consumption records within the specified time periods were used to calculate seven weakly correlated features. Projection methods such as Principal Component Analysis (PCA) were also used to concisely represent a consumer's load curve [113].

In the context of the energy consumption segmentation application, a feature selection approach based on genetic algorithms has been utilized effectively in [112] to reduce the high dimensionality of smart meter data and improve the clustering performance of k-means. In general, several non-exhaustive search methods were conducted to perform feature selection, such as sequential forward search, backward search, floating search, beam search, bidirectional search, and genetic search [105–107,129]. However, and more recently, several studies have approached the problem of finding the optimal set of features as an optimization problem within the context of mixture-based clustering in several real-life applications [15,63]; thus, achieving feature selection with minimal computation expense.

For the purpose of finding the true number of energy consumer groups, diverse methods were followed to accomplish this task. Several studies have approached the solution to this issue via the use of diverse clustering evaluation metrics. Different clustering scenarios are evaluated, and the number of consumption profiles corresponding to the best clustering scenario is chosen to be the optimal number of energy consumption profiles [130, 131]. Similarly, an entropy-based model performance evaluation index was utilized for finding the optimal number of clusters in time series data [112]. On the other hand, the probabilistic model selection methods have proven to be a solution that is invariant to the dataset variable types. In several studies, the *Bayesian Information Criterion* (BIC) and the *Akaike Information Criterion* (AIC) have been used to select the optimal number of energy consumer groups [73, 132]. A study has previously concluded that the performance of the AIC and BIC criteria are affected by several factors that are relevant to the model and the data [133]. Using a smaller and less representative training dataset, BIC tends to select an overly simple model. Additionally, AIC penalizes complex models less. Thus, in specific cases, AIC emphasizes the model's performance using the training dataset and selects more complex models. The *Minimum Message Length* (MML) is another probabilistic model selection measure that is well known to outperform the BIC and the AIC model selection criteria [134–136]. MML coupled with the feature weighing mixture model [62] encourages irrelevant feature weights to degrade to zero, simultaneously performing model selection and feature selection. Thus, avoiding an exhaustive search to find the optimal set of features.

The current energy consumer segmentation approach distinguishes itself from previous works by effectively modelling different representations of smart meter records, inferring the true number of consumer groups and finding the optimal set of features in a single optimization process. The rest of the chapter is organized as follows: in Section 2.3, we describe the proposed *Feature Selection model based on the Bounded Asymmetric Generalized Gaussian Mixture* (FSBAGGMM). Section

2.3 explains how the mixture model's parameters are estimated and how the MML's objective function is derived for our specific case. Section 2.4 exhibits the experimental results in the context of the household energy consumption segmentation by comparing the performance of our proposed algorithm against several state-of-the-art clustering algorithms. Finally, we conclude our research in Section 2.5.

2.2 The unsupervised BAGGMM-based feature selection model

Mixture models are a powerful approach to modelling incomplete data. The observations in this chapter are represented as a set of vectors $\mathcal{X} = \{\vec{X_1}, \vec{X_2}, \vec{X_3}, \dots, \vec{X_N}\}$, $\vec{X_i} \in \mathbb{R}^D$, $i \in \{1, 2, 3, \dots, N\}$. We aim to model data in $\mathcal{X}$ using a mixture model with M components where $M \geq 1$. It is possible to state that the D-dimensional random variable $\vec{X_i} = (X_{i1}, X_{i2}, \dots, X_{iD})$ is sampled from a M component mixture model if it's probability density function can be written as follows:

$$p(\vec{X_i}|\Theta) = \sum_{k=1}^{M} p(\vec{X_i}|\theta_k) p_k \tag{1}$$

where Θ represents the set of parameters of the M-component mixture model. The term p_k represents the mixing proportion of the component k, by definition, p_k is positive and $\sum_{k=1}^{M} p_k = 1$. The likelihood function gives the joint distribution for all the observations:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{k=1}^{M} p(\vec{X_i}|\theta_k) p_k \tag{2}$$

In order to define the complete data likelihood, an M dimensional vector of unobserved variables is defined, and it is denoted by $\vec{Z_i}$. For each observation i, the unobserved binary vector is assigned with 0's except at the k'th position where the cluster is primarily responsible. The complete data likelihood is defined as such:

$$p(\mathcal{X}, Z|\Theta) = \prod_{i=1}^{N} \prod_{k=1}^{M} \left(p(\vec{X_i}|\theta_k) p_k \right)^{Z_{ik}} \tag{3}$$

where $Z = \{\vec{Z}_1, \dots, \vec{Z}_N\}$.The features in Equation 2 are considered to be of equal importance. However, in the context of a real application, the estimation of the feature weights is an effective approach to better model data [113, 114]. The integration of the feature selection approach within the mixture model involves considering that the irrelevant features are modelled with a background Gaussian distribution as in [62]. In this chapter, feature weights are estimated for all the mixture components. Therefore the background Gaussian distribution has a single set of parameters $\vec{\beta} = \{\vec{\eta}, \vec{\delta}\}$. Where $\vec{\eta}$ represents the vector of means for all the data dimensions and $\vec{\delta}$ represents the standard deviations vector. Thus, we are proposing to rewrite Equation 2 to adopt feature relevancy as follows:

$$p(\vec{X}_i|\Theta, \vec{\beta}, \vec{\varphi}) = \sum_{k=1}^{M} p_j \prod_{d=1}^{D} p(X_{id}|\theta_{kd})^{\varphi_d} p(X_{id}|\beta_d)^{1-\varphi_d} \tag{4}$$

where $\vec{\beta} = \{(\eta_1, \delta_1), \dots, (\eta_D, \delta_D)\}$. The unobserved binary vector $\vec{\varphi} = (\varphi_1, ..., \varphi_D)$ indicates the relevancy of each feature. By assuming that the elements within vector $\vec{\varphi}$ are mutually exclusive and independent of the component label Z, thus:

$$p(\vec{X}_i, \vec{\varphi}) = p(\vec{X}_i|\vec{\varphi})p(\vec{\varphi}) = \sum_{k=1}^{M} p_k \prod_{d=1}^{D} \left(\omega_d p(X_{id}|\theta_{kd})\right)^{\varphi_d} \times \left((1-\omega_d)p(X_{id}|\beta_d)\right)^{1-\varphi_d} \tag{5}$$

After the marginalization over φ, the obtained mixture model is formalized as follows:

$$p(\vec{X}_i|\Theta_M) = \sum_{k=1}^{M} p_k \prod_{d=1}^{D} \left[\omega_d p(X_{id}|\theta_{kd}) + (1-\omega_d)p(X_{id}|\beta_d)\right] \tag{6}$$

where $\Theta_M = [\Theta, \vec{\omega}, \vec{\beta}]$ is the complete set of parameters that define the proposed mixture model. The vector $\vec{\omega} = (\omega_1, \dots, \omega_D)$ quantifies the feature importance with a set of weights where $\omega_d = p(\varphi_d = 1)$. Thus, Equation 6 represents the probability density function that is assumed to generate the data. The foreground distribution or the mixture base distribution $p(X_{id}|\theta_{kd})$ models the relevant attributes of each latent class in the data. Several distributions have been proposed for feature selection in the context of mixture models, such as the AGD [15] and the AGGD [63]. However, these distributions are unbounded with a support region that extends across the set of real numbers. Real-life datasets are mostly digitized and have bounded support [21]. Therefore, we propose the

Bounded Asymmetric Generalized Gaussian Distribution (BAGGD) to model the relevant features of each component in the mixture. The BAGGD distribution generalizes several different distribution classes such as the impulsive, the Laplacian, the Gaussian, and the uniform distribution to fit different shapes of observed bounded support, asymmetric and non-Gaussian data. In order to define the bounded distribution proposed in this chapter, the bounded support region τ_{kd} in $\mathbb{R}$ for each component is first defined for the following indicator function:

$$H(X_{id}|k) = \begin{cases} 1 & X_{id} \in \tau_{kd} \\ 0 & Otherwise \end{cases} \tag{7}$$

The bounded asymmetric generalized Gaussian probability density function for each D-dimensional data point is defined following [137] as follows:

$$p(\vec{X}_i|\theta_k) = \prod_{d=1}^{D} \frac{\Psi(X_{id}|\theta_{kd})H(X_{id}|k)}{\int_{\partial_k} \Psi(X_{id}|\theta_{kd})dX} \tag{8}$$

The unbounded distribution $p(X_{id}|\theta_{kd})$ is the *Asymmetric Generalized Gaussian Distribution* (AGGD). The symmetric and asymmetric generalized Gaussian distributions are defined in Equations 9 and 10, respectively.

$$g(X_{id}|\mu_{kd}, \sigma_{kd}, \lambda_{kd}) = \frac{\lambda_{kd}\left[\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right]^{1/2}}{2\sigma_{kd}\Gamma(1/\lambda_{kd})} exp\left[- A(\lambda_{kd})\left|\frac{X_{id} - \mu_{kd}}{\sigma_{kd}}\right|^{\lambda_{kd}}\right] \tag{9}$$

$$\begin{aligned} \Psi(X_{id}|\theta_{kd}) = \begin{cases} g_1(X_{id}|\theta_{kd}) & x < \mu_{kd} \\ \\ g_2(X_{id}|\theta_{kd}) & x \geq \mu_{kd} \end{cases} &= \frac{\lambda_{kd}\left[\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right]^{1/2}}{(\sigma_{l_{kd}} + \sigma_{r_{kd}})\Gamma(1/\lambda_{kd})} \\ &\times \begin{cases} exp\left[- A(\lambda_{kd})\left(\frac{\mu_{kd} - X_{id}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}}\right] & X_{id} < \mu_{kd} \\ exp\left[- A(\lambda_{kd})\left(\frac{X_{id} - \mu_{kd}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}}\right] & X_{id} \geq \mu_{kd} \end{cases} \end{aligned} \tag{10}$$

where $A(\lambda_{kd}) = \left[\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right]^{\lambda_{kd}/2}$; $\theta_{kd} = [\mu_{kd}, \sigma_{l_{kd}}, \sigma_{r_{kd}}, \lambda_{kd}]$ represents the set of parameters that defines the AGGD for each mixture component. μ_{kd}, $\sigma_{l_{kd}}$, $\sigma_{r_{kd}}$ and λ_{kd} denote the mean, the left standard deviation, the right standard deviation and the shape parameter of the AGGD, respectively. The shape parameter controls the distribution's tails. The larger its value, the flatter the distribution at the mean; the smaller it is, the more peaked the distribution at the mean. The right and left variance combination allows the probability density function to be asymmetric or non-asymmetric. Thus, The proposed mixture model would consider the different shapes, asymmetry, and bounded support region of the smart meter data. Bounded distribution generalizes to all its special cases, including the bounded variants [21]. Thus, our proposed FSBAGGMM generalizes to a wide range of mixture models, including the bounded variants, as shown in Table 2.1. Additionally, We will demonstrate in section 2.4 how the proposed FSBAGGMM can generalize feature selection model based on the asymmetric generalized Gaussian mixture in addition to several specific mixture models in terms of modelling smart meter data.

Special Case	**Required Change in FSBAGGMM Parameters**
Feature Selection model based on the Asymmetric Generalized Gaussian Mixture (FSAGGMM) [63]	$H(X_{id}\|k) = 1$
Feature Selection model based on the Bounded Asymmetric Gaussian Mixture (FSBAGMM)	$\lambda_{kd} = 2$
Feature Selection model based on the Asymmetric Gaussian Mixture (FSAGMM) [15]	$H(X_{id}\|k) = 1, \lambda_{kd} = 2$
Feature Selection model based on the Bounded Generalized Gaussian Mixture (FSBGGMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}$
Feature Selection model based on the Generalized Gaussian Mixture (FSGGMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}, H(X_{id}\|k)=1}$
Feature Selection model based on the Bounded Gaussian Mixture (FSBGMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 2$
Feature Selection model based on the Gaussian Mixture (FSGMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 2, H(X_{id}\|k) = 1$
Feature Selection model based on the Bounded Laplace Mixture (FSBLMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 1$
Feature Selection model based on the Laplace Mixture (FSLMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 1, H(X_{id}\|k) = 1$
Asymmetric Generalized Gaussian Mixture Model (AGGMM) [125]	$H(X_{id}\|k) = 1, \omega_d = 1$
Bounded Asymmetric Gaussian Mixture Model (BAGMM)	$\lambda_{kd} = 2, \omega_d = 1$
Asymmetric Gaussian Mixture Model (AGMM) [138]	$H(X_{id}\|k) = 1, \lambda_{kd} = 2, \omega_d = 1$
Bounded Generalized Gaussian Mixture Model (BGGMM) [21]	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \omega_d = 1$
Generalized Gaussian Mixture Model (GGMM) [40]	$\sigma_{r_{kd}} = \sigma_{l_{kd}, H(X_{id}\|k)=1}, \omega_d = 1$
Bounded Gaussian Mixture Model (BGMM) [22]	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 2, \omega_d = 1$
Gaussian Mixture Model (GMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 2, H(X_{id}\|k) = 1, \omega_d = 1$
Bounded Laplace Mixture Model (BLMM) [44]	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 1, \omega_d = 1$
Laplace Mixture Model (LMM)	$\sigma_{r_{kd}} = \sigma_{l_{kd}}, \lambda_{kd} = 1, H(X_{id}\|k) = 1, \omega_d = 1$

Table 2.1: FSBAGGMM special cases

2.3 Model Parameter Estimation and Selection

In this section, we will explain how the feature weights and the mixture model parameters are estimated for modelling the training data in addition to the model selection criterion. We propose an approach to reveal the valid number of intrinsic groups within a dataset using MML and estimate the proposed model's parameters using EM.

2.3.1 Parameter estimation using the EM algorithm

The mixture model's parameters are optimized in parallel with the features' weights in each iteration using the EM algorithm. The iterations of the EM algorithm produce a sequence of models with a non-decreasing log-likelihood. The parameters are optimized to achieve the maximum log-likelihood, and the log-likelihood function is expressed as follows:

$$\begin{aligned}\mathcal{L}(\mathcal{X},\Theta_M,Z,\varphi) = \sum_{i,k} p(Z_i = k|\vec{X_i})\log p_k + \sum_{i,k}\sum_{d}\sum_{\varphi_d=0}^{1} p(Z_i = k,\varphi|\vec{X_i}) \\ \times \Big(\varphi_d(\log(p(X_{id}|\theta_{kd}) + \log w_d) + (1-\varphi_d)(\log p(X_{id}|\beta_d) + \log(1-\omega_d))\Big)\end{aligned} \tag{11}$$

The EM algorithm has made the optimization process for mixture models feasible through an iterative process using Equation 11 instead of Equation 2. The conditional expected values $\gamma(Z_{jh})$ and $\hat{\omega_d}$ are given by Equations 12 and 13.

$$p(Z_i = k|\vec{X}_i,\Theta_M) = \gamma(Z_{ik}) = \frac{p_k \prod_{d=1}^{D}\zeta_{i,k,d}}{\sum_{j=1}^{K} p_j \prod_{d=1}^{D}\zeta_{i,j,d}} \tag{12}$$

$$\hat{\omega_d} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p(X_{id}|\theta_{jd})}{\zeta_{i,j,d}}\gamma(Z_{ij})}{N} \tag{13}$$

where $\zeta_{i,k,d} = \big[\omega_d p(X_{id}|\theta_{kd}) + (1-\omega_d)p(X_{id}|\beta_d)\big]$. The EM algorithm consists of a loop over two steps: the E-step and the M-step; they are performed repetitively until convergence. In the E-step, Equation 12 is evaluated using either the initial parameters or the parameters estimated in the M-step. In the M-Step, the parameters of the next model in the sequence are estimated. Each estimated model in the sequence represents a better approximation of the distribution of the smart meter data. Due to the complicated nature of the BAGGD function, the gradient of the Log-Likelihood function (Equation 11) with respect to each one of the parameters was non-linear, and a closed-form solution was not obtained; therefore, for these parameters, we used the Newton-Raphson method to approximate the update values as demonstrated in the equations below. The partial derivatives obtained with respect to each of the parameters can be found in Appendix A.1. Thus, the M-step is implemented using the following equations:

$$p_k = p(Z_k = 1) = \frac{\sum_{i=1}^{N} p(k|\vec{X}_i, \Theta_M)}{N} \tag{14}$$

$$\hat{\mu_{kd}} = \mu_{kd} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \mu_{kd}^2}\right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \mu_{kd}}\right)\right] \tag{15}$$

$$\hat{\sigma_{l_{kd}}} = \sigma_{l_{kd}} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{l_{kd}}^2}\right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{l_{kd}}}\right)\right] \tag{16}$$

$$\hat{\sigma_{r_{kd}}} = \sigma_{r_{kd}} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{r_{kd}}^2}\right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{r_{kd}}}\right)\right] \tag{17}$$

$$\hat{\lambda_{kd}} = \lambda_{kd} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \lambda_{kd}^2}\right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \lambda_{kd}}\right)\right] \tag{18}$$

$$\hat{\eta_d} = \frac{\sum_{i=1}^{N} \left[\frac{(1-\omega_d)p(X_{id}|\beta_d)}{\zeta_{i,k,d}} \gamma(Z_{ik})\right] x_{id}}{\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{(1-\omega_d)p(X_{id}|\beta_d)}{\zeta_{i,j,d}} \gamma(Z_{ij})} \tag{19}$$

$$\hat{\delta_d}^2 = \frac{\sum_{i=1}^{N} \left[\frac{(1-\omega_d)p(X_{id}|\beta_d)}{\zeta_{i,k,d}} \gamma(Z_{ik})\right] (x_{id} - \eta_d)^2}{\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{(1-\omega_d)p(X_{id}|\beta_d)}{\zeta_{i,j,d}} \gamma(Z_{ij})} \tag{20}$$

2.3.2 Model Selection

Model selection involves selecting the best set of parameters that model the smart meter data. Among several candidate models, the model with the maximum log-likelihood may achieve the best fit to the data; however, it is not guaranteed to perform well on unseen data. In other words, model evaluation based on the log-likelihood exclusively could be misleading. In this section, we develop a model selection criterion to infer the true number of consumption profiles within a dataset in an unsupervised manner. The MML criterion [139, 140] is an information theory-based model selection method; it selects the best model among a list of candidate statistical models based on its capability of compressing a message containing the data. According to the MML criterion, the best model minimizes a message that consists of two parts; the first part encodes the model using prior knowledge about the model exclusively, and the second part encodes the data using the model. Given a list of candidate models, the following function is minimized to obtain the true number of intrinsic groups within the data:

$$\text{MessLens} \approx -\log p(\Theta_M) + \frac{c}{2}(1 + \log \rho_c) + \frac{1}{2}\log |I(\Theta_M)| - \log p(\mathcal{X}|\Theta_M) \tag{21}$$

In equation 21, the prior distribution is represented by $p(\Theta_M)$, the determinant of the Fisher information matrix is represented by $|I(\Theta_M)|$, the model's likelihood is represented by $p(\mathcal{X}|\Theta_M)$. The constant c is the total number of parameters; in this case, it is calculated as such $c = M + D + 4DM + 2D, c \geq 1$. The term $\rho_c \in \mathbb{R}^c$ represents the optimal quantization lattice constant [141]; the value of the constant is approximated with $\rho_c = \frac{1}{12}$ as the value of c changes across the list of candidate models [142].The independence of the different groups of parameters has been considered in this chapter; which allows the factorization of the prior distribution and Fisher information matrix in equation 21. Additionally, we approximate the determinant of the Fisher information matrix using the complete likelihood, and we consider the uninformative Jeffrey's prior for the distribution of each group of parameters. Hence, in our case, the MML optimization objective function is calculated as such:

$$\begin{aligned} \text{MessLens} \approx & \frac{c}{2}(1 + \log \rho_c) + \frac{c}{2}(\log N) + 2M\sum_{d=1}^{D}\log \omega_d \\ & + 2d\sum_{k=1}^{M}\log p_k + \sum_{d=1}^{D}\log(1-\omega_d) - \log p(\mathcal{X}|\Theta_M) \end{aligned} \tag{22}$$

Equation 22 is minimized with respect to the several constraints [62], which are listed as follows: $0 < p_k \leq 1$, $0 \leq \omega_d \leq 1$ and $\sum_{j=1}^{M} p_j = 1$. In the context of this model selection criterion, since we are estimating feature weights using the EM algorithm, Equation 23 and 24 are utilized alternatively to approximate the parameters $\hat{p_k}$ and $\hat{\omega_d}$ respectively and as follows:

$$\hat{p_k} = \frac{\max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M}\gamma(Z_{ij}) - 2D, 0\Big)}{\sum_{j=1}^{M}\max\Big(\sum_{i=1}^{N}\gamma(Z_{ij}) - 2D, 0\Big)} \tag{23}$$

$$\hat{\omega_d} = \frac{\max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p(X_{id}|\theta_{jd})}{\zeta_{i,j,d}}\gamma(Z_{ij}) - 2M, 0\Big)}{T} \tag{24}$$

$$T = \max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p(X_{id}|\theta_{jd})}{\zeta_{i,j,d}}\gamma(Z_{ij}) - 2M, 0\Big) + \max\Big(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\beta_d)}{\zeta_{i,j,d}}\gamma(Z_{ij}) - 1, 0\Big) \tag{25}$$

The algorithm of model selection and model parameter estimation

Algorithm 1 describes how to perform model selection and feature selection using the MML criterion and model parameter estimation using the EM algorithm.

2.4 Experimental Results

In this section, we will validate the performance of the MML model selection criterion and the proposed FSBAGGMM using two synthetic and real-life smart meter datasets within the application of household energy consumption segmentation. The first real-life dataset was recorded by the *Commission for Energy Regulation* (CER) and made accessible for researchers by the *Irish Social Science Data Archive* (ISSDA) [89]. The dataset consists of smart meter records gathered from more than 6000 Irish energy consumers from July 14, 2009, to December 31, 2010. The energy consumption is recorded in kWh with an interval of half an hour. This dataset has two types of energy consumers: residential and small to medium enterprises. As stated earlier, we are interested in analyzing the energy consumption of residential energy consumers only. Therefore, 3639 Irish residential energy consumers remain for analysis after data cleaning. Each residential consumer is assigned six different tariffs (E, A, D, C, B, and W). The second real-life smart meter dataset consists of smart meter records collected from 5567 residential homes in London. The data was collected

Algorithm 1 Unsupervised FSBAGGMM

```
While $M < M_{max}$ do
    Initialize $\Theta_M$
        A  K-Means clustering results are used to initialize the parameters
           $(\pi_1, \ldots, \pi_M, \vec{\mu_1}, \ldots, \vec{\mu_M}, \vec{\sigma_{l_1}}, \ldots, \vec{\sigma_{l_M}}, \vec{\sigma_{r_1}}, \ldots, \sigma_{r_M}, \lambda_1, \vec{\ldots}, \lambda_M)$.
        B  For each cluster $k$, each element of the parameter vector $\vec{\lambda_k}$ is set to the value 2.
        C  Initialize the background Gaussian distribution parameter set $\vec{\beta}$ using the following Equations for all the dimensions, where $d \in \{1, \ldots, D\}$:

               $\eta_d = \frac{1}{N} \sum_{i=1}^{N} X_{id}$    (26)

               $\delta_d^2 = \frac{1}{N} \sum_{i=1}^{N} (X_{id} - \eta_d)^2$    (27)

    Implement the E-step.
        (1)  For each cluster $k$, compute the bounded support region $\vec{\tau_k} = (\tau_1, \ldots, \tau_D)$.
        (2)  Evaluate equation 12.
               • if $\omega_d = 0$ Then $p(X_{id}|\theta_{kd}) = 0$
               • if $\omega_d = 1$ Then $p(X_{id}|\beta_d) = 0$
    Implement the M-Step using Equations 15 through 20, 23 and 24.
    if $p(\mathcal{X}|\Theta)^{\iota+1} - p(\mathcal{X}|\Theta)^{\iota} < \epsilon$ then
        • Calculate the message length using Equation 22.
```

by the *UK Power Networks* that is led by the *Low Carbon London project* between November 2011 and February 2014. The energy consumption is recorded in kWh with an interval of half an hour UKPN2013. After data cleaning, observations of 3891 household energy consumers within the year 2013 are used to analyze this experiment. The residential energy consumers in this dataset are subjected to two types of tariffs; the first type is the *Dynamic Time of Use* (ToU), where the energy consumption prices vary as follows: High (67.20 pence/kWh), Low (3.99 pence/kWh) or normal (11.76 pence/kWh), the second type is the *Standard* (std), where the consumers paid a flat rate of 14.228 pence/kWh. Additionally, the energy consumers in this dataset belong to five different geo-demographic groups.

The application used in this chapter aims to segment energy consumers given their load curve. We use characteristic load profiles to find the optimal number of energy consumption groups with similar consumption patterns and determine the cluster membership of every load curve given in the training dataset. Utility companies can use accurate energy consumer-type identification to make correct decisions regarding the investments in load-shifting campaigns to prevent over or under-dimensioning linked to peak energy demand. Several performance evaluation metrics were introduced in [131]. They are defined as follows:

DI [143]: Dunn's index is a model performance evaluation metric that is calculated using the minimum ratio between the closest distance of two observations of different clusters and the largest distance between two observations in the same cluster. This index is maximized for best clustering, and it is defined as follows:

$$\text{DI} = \frac{\min_{A \in M}\big\{\min_{B \in M, B \neq A}\{\phi(A,B)\}\big\}}{\max_{A \in M}\{\Pi(A)\}} \tag{28}$$

$$\phi(A,B) = \min_{\vec{X}_i \in A, \vec{Y}_j \in B}\big\{d(\vec{X}_i, \vec{Y}_j)\big\} \tag{29}$$

$$\Pi(A) = \max_{\vec{X}_i, \vec{X}_j \in A}\big\{d(\vec{X}_i, \vec{X}_j)\big\} \tag{30}$$

where d denotes the distance or the similarity function, $\phi(A, B)$ denotes the minimum distances between two observations that each belong to either cluster A or B, and M denotes the set of clusters.

EoE [112]: The *Entropy of Eigenvalue* is an entropy-based clustering performance measure; it is obtained from the eigenvalue analysis of the correlation matrix calculated using raw smart meter data. The index is calculated using the correlation between representative time series of different clusters and the correlation between different time series within each cluster. The EoE index is calculated using the following equation:

$$\text{EoE} = \frac{SM_B}{\sum_k^K \frac{N_k}{N} SM_{wk}} \tag{31}$$

The SM similarity is a normalized average information measure; the larger it is, the greater the similarity. The term SM_b represents the normalized entropy of eigenvalues obtained from the correlation matrix between different clusters, and SM_{wk} represents the normalized entropy of eigenvalues obtained from the correlation matrix between time series in each cluster k. In an ideal clustering, EoE is a small value consisting of high similarity between time series within each cluster and low similarity between representative time series of different clusters.

S [144]: The Silhouette score is a model evaluation measure that is concerned with calculating a score for each observation in the training dataset. The measure calculates the overall evaluation by computing the average score for all the dataset observations. The metric is maximized for better clustering and is defined in the following equation:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}} \tag{32}$$

where $a(x_i)$ represents the average dissimilarity of the data point x_i to all the other data points within the same cluster. $b(x_i)$ represents the minimum average dissimilarity of data point x_i to data points existing in a cluster different from the data point's cluster.

CH [145]: The *Calinski-Harabasz* is a model performance evaluation index; the measure calculates the ratio between the inter-cluster variance and the intra-cluster variance. This measure is maximized for better clustering and is defined as follows:

$$\text{CH} = \frac{N-K}{K-1}\frac{\sum_{k=1}^{K}\left(N_k d(c_k,\bar{c})\right)}{\sum_{k=1}^{K}\sum_{i=1}^{N_k} d(\vec{X}_i, c_k)} \tag{33}$$

where N_k is the number of observations predicted to belong to cluster k, c_k denotes the centroid of class k, $\bar{c}$ denotes the global centroid of all the clusters, and d denotes the distance or the similarity function.

DB [146]: The *Davies–Bouldin* index is a model performance evaluation measure; it calculates the ratio of intra-cluster distances to inter-cluster distances for each possible pair of clusters. The maximum ratio calculated for each pair of clusters is considered in a summation. The summation result is divided by the total number of clusters to obtain the metric's value. This measure is minimized for better clustering, and it is defined as follows:

$$\text{DB} = \frac{1}{k}\sum_{A\in M} \max_{B\in M, B\neq A}\left\{\frac{O(A)+O(B)}{d(c_A,c_B)}\right\} \tag{34}$$

$$O(A) = \frac{1}{\varrho(A)}\sum_{\vec{X}_i \in A} d(\vec{X}_i, c_A) \tag{35}$$

where $\varrho(A)$ denotes the cardinality of cluster A, k denotes the number of components enforced by the mixture model, M denotes the set of clusters, c_A denotes the centroid of class A, and d denotes the distance or the similarity function. M has k elements

GOF [147]: The *Goodness of Fit* statistic value measures the model's fitting accuracy, and it is calculated as follows:

$$\text{GOF} = \sum_{i=1}^{N}\frac{(\Upsilon(\vec{X}_i) - \Omega(\vec{X}_i))^2}{\Omega(\vec{X}_i)} \tag{36}$$

where $\Upsilon(\vec{X}_i)$ and $\Omega(\vec{X}_i)$ represent the empirical and the expected frequencies of the observation

$\vec{X}_i$ respectively. The indices ACC, TPR, PPV, TNR, NPV, FPR, FNR, and FDR, represent average accuracy, average true positive rate, positive predictive value, true negative rate, negative predictive value, false-positive rate, false-negative rate, and false discovery rate, respectively. They are defined as follows:

$$\mathrm{TPR} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{37}$$

$$\mathrm{TNR} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{TN}_k}{\mathrm{TN}_k + \mathrm{FP}_k} \tag{38}$$

$$\mathrm{PPV} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{39}$$

$$\mathrm{NPV} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{TN}_k}{\mathrm{TN}_k + \mathrm{FN}_k} \tag{40}$$

$$\mathrm{FPR} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{FP}_k}{\mathrm{FP}_k + \mathrm{TN}_k} \tag{41}$$

$$\mathrm{FNR} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{FN}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{42}$$

$$\mathrm{FDR} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{FP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{43}$$

$$\mathrm{ACC} = \frac{1}{M}\sum_{k=1}^{M}\frac{\mathrm{TP}_k + \mathrm{TN}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k + \mathrm{TN}_k} \tag{44}$$

where TP_k, FP_k, TN_k, and FN_k denote the number of true positives, false positives, true negatives and false negatives respectively for the cluster k. In order to compute the metrics explained in Equations 37 to 44, cluster k labels are considered a positive class, and all the remaining cluster labels are considered a negative class. MCC represents the *Mathiews Correlation Coefficient* evaluation metric [148].

AIC and BIC are probabilistic model selection methods [149] that attempt to select the model with the best performance while taking into consideration its complexity (by adding a complexity-related penalty). Unlike probabilistic model selection criteria, performance metrics select models with disregard for their complexity. The distinct probabilistic model selection criteria used in this chapter originate from different fields of study. AIC is derived from the frequentist framework, while BIC is derived from Bayesian probability and inference. Compared to BIC, AIC emphasizes the model performance and penalizes complex models less, making it prone to select overfitted models. In comparison to AIC, BIC attempts to penalize candidate models more for their complexity. The AIC and BIC model selection criteria statistics for each candidate model are computed as follows:

$$BIC = 2\log(L(\Theta)) + \kappa \log(N) \tag{45}$$

$$AIC = \frac{-2}{N}\log(L(\Theta)) + 2 * \frac{\kappa}{N} \tag{46}$$

where $L(\Theta)$ is the likelihood function estimate given a set of parameters Θ , κ represents the number of free parameters, and N represents the number of observations. As N approaches infinity, the BIC criterion is more likely to select the candidate model with the true number of intrinsic groups. The candidate model with the lowest AIC and BIC are selected for both model selection criteria.

In the upcoming sections, The performance of the proposed model is compared to specific mixture models such as the BAGGMM, the AGGMM, and the FSAGGMM. Model selection using the proposed model is performed using the MML model selection criterion and compared against specific model selection methods such as BIC, AIC, and model selection methods using performance measures such as the *Dunn's Index* (DI), and the *Entropy of Eigenvalue* (EOE).

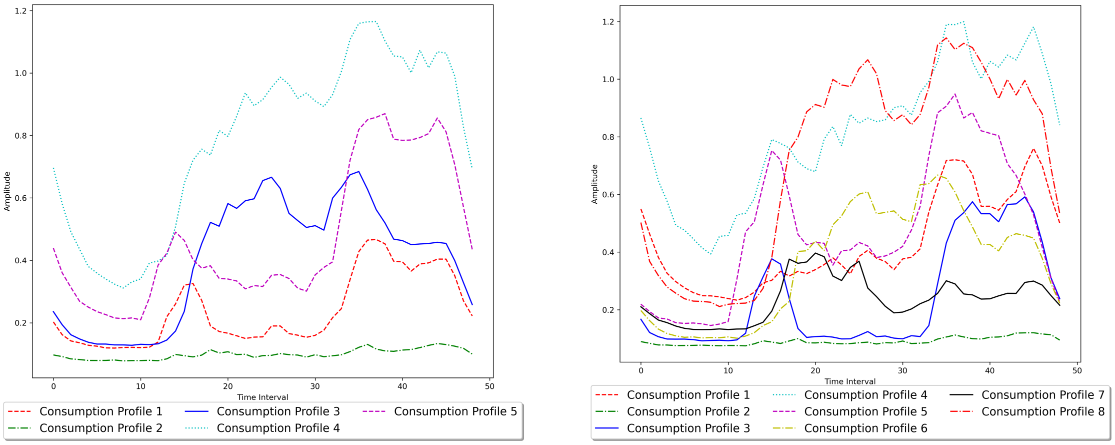
2.4.1 Synthetic Data

As a first stage, synthetic datasets are used to validate the proposed mixture model and its model selection method. We propose using a 49-dimensional dataset, which imitates a real-life smart meter records dataset by representing each energy consumer with a load curve. In order to generate the synthetic datasets used in this chapter, the following steps must be followed:

(1) For each energy consumer in the real-life dataset, only the first 49 smart meter records are considered.

(2) The Gaussian mixture model is used to cluster the data into a specific number of clusters. The mean of each cluster is considered a consumption profile.

(3) Each consumption profile inferred from the previous step is summed with instances generated by a Gaussian white noise using five different sets of parameters to form the observations of the synthetic dataset.



(a) First synthetic dataset

(b) Second synthetic dataset

Figure 2.2: Consumption profiles used to generate the synthetic datasets

In other words, the origin of each cluster of observations within the synthetic datasets used in this chapter is an actual energy consumption profile concluded from a real dataset. The first

Gaussian White Noise Parameters	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5
μ=0.001; σ=0.2	378	370	379	371	382
μ=0.01; σ=0.2	349	364	356	356	355
μ=0.1; σ=0.2	352	360	361	359	348
μ=0.05; σ=0.3	354	358	359	356	353
μ=0.01; σ=0.3	365	353	357	350	355

Table 2.2: Count of observations generated for the first synthetic dataset

dataset consists of five clusters. The five real-life consumption profiles used to generate the first dataset are demonstrated in Figure 2.2a. The count of the observations generated for each energy consumption profile using the distinct Gaussian white noise parameters is shown in Table 2.2. As an illustrative example of the data generation process, 378 observations of the first dataset are generated by summing the white noise vector generated using the parameter set ($\mu = 0.001; \sigma = 0.2$) of the multivariate Gaussian white noise with the vector of ”Consumption Profile 1”.

Our model selection approach successfully infers the correct number of components within this dataset, as demonstrated in Table 2.3. MML outperforms specific model selection methods using

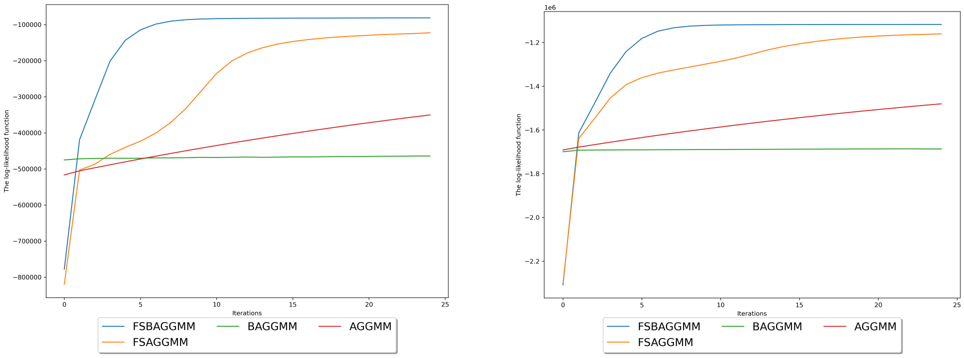
	FSBAGGMM
Model Selection Method	
BIC	7
AIC	7
DI	4
MML	5
EoE	5
Ground Truth	5

Table 2.3: Identified optimal number of clusters using the first synthetic dataset

	BAGGMM + FW
Model Selection Method	
BIC	6
AIC	6
DI	6
MML	8
EoE	8
Ground Truth	8

Table 2.4: Identified optimal number of clusters using the second synthetic dataset

the clustering results obtained from each instance of our proposed model.



(a) First synthetic dataset

(b) Second synthetic dataset

Figure 2.3: Mixture model's log-likelihood functions demonstration during the clustering of the Synthetic Datasets

Figure 2.3a demonstrates the maximum log-likelihood achieved by clustering the data using the proposed model in comparison with specific mixture models. The clustering results of our proposed model are evaluated using several performance measures and compared against the clustering performances of specific mixture models as shown in Tables 2.6, and 2.7. The proposed model achieves the best fit of the training data by scoring the best performance according to all the performance metrics used in this experiment and by reaching the highest log-likelihood.

The second dataset consists of eight clusters. The eight real-life consumption profiles used to

Gaussian White Noise Parameters	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5	Profile 6	Profile 7	Profile 8
μ=0.001; σ=0.2	445	448	450	444	449	447	442	455
μ=0.01; σ=0.2	442	449	448	448	448	452	445	448
μ=0.1; σ=0.2	442	452	455	449	447	443	447	445
μ=0.05; σ=0.3	445	448	444	451	453	447	442	450
μ=0.01; σ=0.3	460	459	458	468	457	455	466	457

Table 2.5: Count of observations generated for the second synthetic dataset

generate this dataset are demonstrated in Figure 2.2b. The count of the observations generated for each energy consumption profile using the distinct Gaussian white noise parameters is shown in Table 2.5. Our model selection approach successfully infers the correct number of components within this dataset, as demonstrated in Table 2.4. MML chooses the proposed model's instance with a component count equal to the ground truth outperforming specific model selection methods used in this comparison. The proposed model fits the data better than all the mixture models used in the comparison by achieving the highest maximum log-likelihood as demonstrated in Figure 2.3b. According to all the performance metrics used in this experiment, the proposed model also outperforms the mixture models selected for the comparison as shown in Tables 2.8, and 2.9.

Performance Index (%)	FSBAGGMM	FSAGGMM	BAGGMM	AGGMM
ACC	95.569	94.338	85.458	82.804
TPR/Recall	88.935	85.836	63.589	56.953
PPV/Precision	89.458	88.149	74.838	70.500
MCC	86.291	82.921	58.170	51.104
F1-Score	88.922	85.844	63.644	57.011
TNR	97.231	96.461	90.906	89.245
NPV	97.263	96.591	92.128	90.942
FPR	2.769	3.539	9.094	10.755
FNR	11.065	14.164	36.411	43.047
FDR	10.542	11.851	25.162	29.500

Table 2.6: Mixture models' clustering performance evaluation using the first synthetic dataset

Performance Index	Optimal Performance Indicator	FSBAGGMM	FSAGGMM	BAGGMM	AGGMM
GOF	Minimum	3870.683	7261.083	16397.633	17765.500
CH	Maximum	2081.868	2046.444	1594.215	1405.947
S	Maximum	0.107	0.100	0.023	-0.016
DB	Minimum	2.549	2.623	2.661	2.503
DI	Maximum	0.224	0.219	0.209	0.209
Xie and Benie Index	Minimum	1.871	1.881	2.446	2.698
Fowlkes Mallows	Maximum	0.799	0.755	0.650	0.648
Log Loss	Minimum	0.625	0.901	9.741	12.138
EOE	Minimum	0.730	0.758	1.022	1.032
Jaccard	Maximum	0.889	0.858	0.636	0.570
ROC AUC	Maximum	0.931	0.912	0.773	0.731
V Measure	Maximum	0.755	0.740	0.660	0.639
Rand	Maximum	0.919	0.899	0.820	0.795
Normalized Mutual Information	Maximum	0.755	0.740	0.660	0.639
Mutual Information	Maximum	1.213	1.181	0.969	0.887
Homogeneity	Maximum	0.754	0.734	0.602	0.551
Adjusted Rand	Maximum	0.749	0.691	0.524	0.497
adjusted mutual info	Maximum	0.755	0.740	0.660	0.639

Table 2.7: Mixture models' clustering performance evaluation using the first synthetic dataset

Performance Index (%)	FSBAGGMM	FSAGGMM	BAGGMM	AGGMM
ACC	91.856	88.746	88.481	87.769
TPR/Recall	67.459	54.969	53.862	51.021
PPV/Precision	66.482	55.753	56.402	54.291
MCC	63.813	50.402	49.908	46.726
F1-Score	67.422	54.983	53.922	51.078
TNR	95.347	93.570	93.418	93.012
NPV	95.456	93.921	93.926	93.528
FPR	4.653	6.430	6.582	6.988
FNR	32.541	45.031	46.138	48.979
FDR	33.518	44.247	43.598	45.709

Table 2.8: Mixture models' clustering performance evaluation using the second synthetic dataset

Performance Index	Optimal Performance Indicator	FSBAGGMM	FSAGGMM	BAGGMM	AGGMM
GOF	Minimum	22539.820	36474.842	50310.225	48011.423
CH	Maximum	2100.955	1766.797	1713.450	1674.616
S	Maximum	0.054	0.001	-0.052	-0.062
DB	Minimum	3.563	4.975	6.767	6.738
DI	Maximum	0.210	0.213	0.208	0.194
Xie and Benie	Minimum	2.883	3.619	3.683	3.784
Fowlkes Mallows	Maximum	0.574	0.486	0.518	0.503
Log Loss	Minimum	3.293	10.287	12.618	13.228
EOE	Minimum	0.620	0.637	0.685	0.675
Jaccard	Maximum	0.674	0.550	0.539	0.511
ROC AUC	Maximum	0.814	0.743	0.737	0.720
V Measure	Maximum	0.644	0.565	0.593	0.586
Rand	Maximum	0.881	0.836	0.831	0.821
Normalized Mutual Information	Maximum	0.644	0.565	0.593	0.586
Mutual Info	Maximum	1.303	1.088	1.114	1.093
Homogeneity	Maximum	0.627	0.523	0.536	0.526
Adjusted Rand	Maximum	0.502	0.384	0.407	0.385
Adjusted Mutual Info	Maximum	0.644	0.565	0.593	0.585

Table 2.9: Mixture models' clustering performance evaluation using the second synthetic dataset

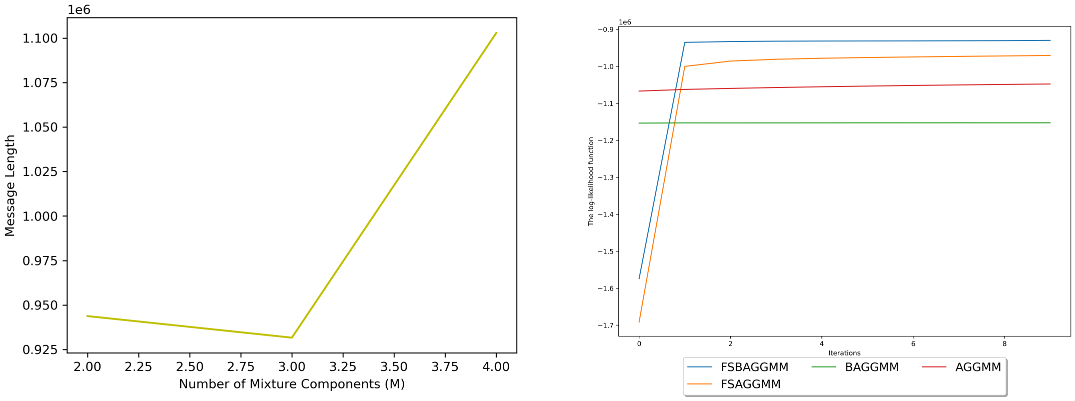
2.4.2 Real-life Smart Meter Data

The commission for energy regulation smart meter data

In this section, we investigate the performance of our proposed model using the first real-life smart meter dataset. As mentioned earlier, the dataset we considered has smart meter records from 3639 Irish energy consumers. Each consumer has 25728 electricity usage readings that are recorded in kilowatt-hours. In order to summarize and preserve the information within the numerous features representing each energy consumer, PCA is used for feature extraction in this experiment. Several datasets with a different number of features were considered within the range between 50 and 250. Due to the low reconstruction error, the dataset with 250 features is favoured for this experiment.

We used the dataset as an input to three different instances of our proposed model. Each instance had a different number of mixture components within the range $M = [2, 4]$. The model selection algorithm concluded that the minimum value calculated using its objective function is obtained while using the model instance with three components, as shown in Figure 2.4a. Table 2.10

demonstrates the optimal number of clusters concluded by each model selection criterion used in comparison with MML. In addition to the fact that our derived model selection criterion has been inferring the correct number of clusters in solid experiments using synthetic data, AIC and BIC also agree that the true number of clusters is three in this experiment.



(a) Selection of the optimal number of mixture components using MML and the proposed model

(b) The log-likelihood functions of the mixture models used in the comparison

Figure 2.4: The mixture Model's performance information during the clustering of the first real-life smart meter data

Figure 2.4b demonstrates the log-likelihood trail for each mixture model used in the comparison within this experiment. As observed, the proposed model has converged to the highest log-likelihood indicating a better fit to the training dataset. The clustering evaluation of the proposed model for the concluded optimal number of clusters is demonstrated in Table 2.11 in comparison with specific mixture models. As demonstrated, our proposed model achieves the best clustering performance according to all the evaluation measures used in the comparison.

	FSBAGGMM
Model Selection Method	
BIC	3
AIC	3
DI	2
MML	3
EoE	4

Table 2.10: Identified optimal number of clusters for the real-life smart meter dataset

As mentioned earlier, we determined the true number of clusters using MML and achieved the best clustering result using our proposed mixture model. Since this is an implementation of a real-life application, it is necessary to analyze the resulting clusters to understand further the energy

Performance Index	Metric's Optimal Value	FSBAGGMM	FSAGGMM	BAGGMM	AGGMM
S	Maximum	0.250	0.216	0.228	0.176
CH	Maximum	7.377	5.824	6.671	5.594
DB	Minimum	16.951	23.832	20.626	24.577
DI	Maximum	0.253	0.238	0.249	0.224
Xie and Benie	Minimum	60.821	72.969	62.157	73.319
EOE	Minimum	1.460	1.764	1.613	1.822

Table 2.11: Mixture models' clustering performance using the real-life smart meter dataset

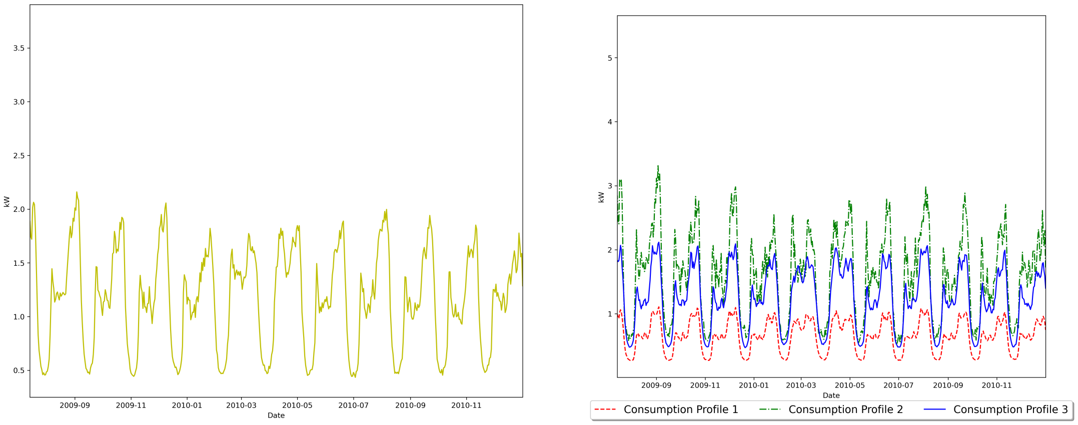
consumption patterns of each consumption trend discovered. Figure 2.5a demonstrates the average power demand of all the energy consumers without clustering. Comparatively, we demonstrate the average power demand of each energy consumer cluster in Figure 2.5b. For all the time intervals available in the dataset, as observed, the responsibility of each energy consumption pattern to the overall average power demand can be determined. The proposed model can determine the consumer's contribution to each consumption profile and which the consumer is mostly following. Table 2.12 demonstrates the ratio of the count of energy consumers in each cluster to the total count of energy consumers in the dataset; the table also demonstrates the consumption responsibility of each consumer cluster to the total average energy consumption in the year 2010. Additionally, the real-life dataset we use in this experiment provides the tariff assigned for each energy consumer. We have discovered that the tariff types are distributed almost identically across the resulting clusters, as shown in Figure 2.6, which indicates that the tariff type does not influence the consumer's electrical usage pattern.

Consumption Profile Cluster	Average Consumption (kWh)	Annual Consumption Responsibility	clusters' proportion
1	6536.770	18.650%	64.600%
2	16117.190	45.980%	1.700%
3	12394.570	35.360%	33.700%

Table 2.12: Consumption profiles statistics for the year 2010

The UK power networks smart meter data

In this section, we validate the performance of our proposed model using the second real-life smart meter data. As mentioned earlier, the dataset we considered in this experiment has smart meter records from 3891 household energy consumers that are located in London. Each consumer has 17520 electricity usage readings that are recorded in kilowatt-hours. In order to summarize the information included in the load curve of each energy consumer, we have extracted nine features.



(a) The average demand of all the energy consumers starting July, 14th 2009 to December,31st 2010

(b) The average demand of the optimal energy consumption clusters from July, 14th 2009 to December,31st 2010

Figure 2.5: Household energy consumption segmentation demonstration of the first real-life smart meter dataset

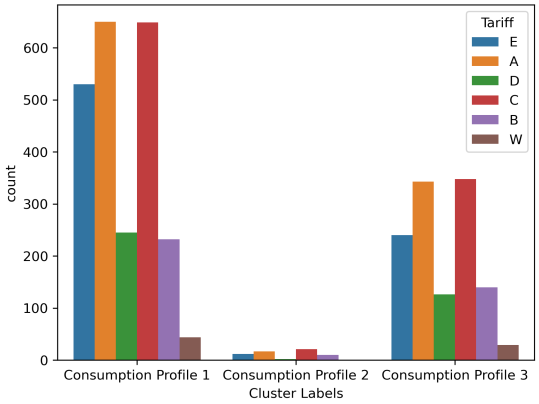


Figure 2.6: Number of energy consumers in each cluster

Following [73], seven features are extracted after the definition of four key time periods, and they are denoted by $t \in \{1, 2, 3, 4\}$. The *Overnight* time period ($t = 1$) is defined between 10:30 PM and 6:30 AM, the *Breakfast* time period ($t = 2$) is defined between 6:30 AM and 9:00 AM, the *Daytime* period ($t = 3$) is defined between 9:00 AM and 3:30 PM, the *Evening* time period ($t = 4$) is defined between 3:30 PM and 10:30 PM. Based on the four previously explained prominent time periods, seven features are extracted from smart meter records to summarize the representation of energy consumers, and they are calculated as follows:

- RAP_t denotes the *Relative Average Power* for time period (t) over the entire year; it is defined as follows:

$$\textbf{RAP}_t = \frac{\text{AP}_t}{\text{DAP}}, t = 1, 2, 3, 4 \tag{47}$$

- the Mean STD denotes the *Mean Relative Standard Deviation* of the average power used over the entire year; it is defined as follows:

$$\textbf{Mean STD} = \frac{1}{4}\sum_{t=1}^{4} \frac{\sigma_t}{\text{AP}_t} \tag{48}$$

- The *seasonal score* is defined as follows:

$$\textbf{Seasonal Score} = \sum_{t=1}^{4} \frac{|AP_t^W - AP_t^S|}{AP_t} \tag{49}$$

- The *Weekend vs Weekday Difference Score* (WD-WE diff. Score) is calculated as follows:

$$\textbf{WD-WE diff. Score} = \sum_{i=1}^{4} \frac{|\text{AP}_t^{\text{WD}} - \text{AP}_t^{\text{WE}}|}{\text{AP}_t} \tag{50}$$

where AP_t, and σ_t represent the average power used by the specific consumer and its corresponding standard deviation in the time period (t) respectively over all the available smart meter records data. DAP represents the average daily power used by the specific consumer throughout the available smart meter data. AP_t^W and AP_t^S represent the average power used by the specific consumer in the time period (t) throughout winter and summer, respectively. AP_t^{WD}, and AP_t^{WE} represent the average power used by the specific consumer in the time period (t) throughout the weekdays and weekends respectively for the available data. Finally, the eighth and the ninth features represent the consumer's tariff and geo-demographic group, respectively.

We have determined the optimal number of clusters for our proposed model using the MML model selection criterion similar to our previous experiments. Among five candidate FSBAGGMM models of mixture components within the range [2,6], the model instance with four components achieved the minimum message length.

	FSBAGGMM
Model Selection Method	
BIC	4
AIC	4
DI	4
MML	4
EoE	2

Table 2.13: Identified optimal number of clusters for the second real-life smart meter dataset

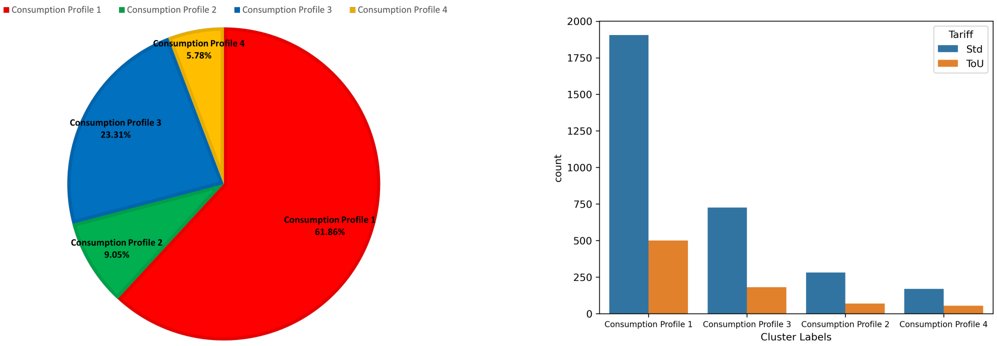
Most of the model selection methods used in the comparison demonstrated in Table 2.13 agree on the optimal number of mixture components. Therefore, the data were clustered into four clusters using our proposed model, and the clustering performance evaluation is compared against specific mixture models. Table 2.14 demonstrates how our proposed mixture model has been able to outperform the different mixture models used in the comparison using six different performance metrics.

	Metric's Optimal Value	**FSBAGGMM**	**FSAGGMM**	**BAGGMM**	**AGGMM**
Performance Index					
S	Maximum	0.319	0.288	0.265	0.189
CH	Maximum	1984.843	1078.837	545.442	243.243
DB	Minimum	1.050	1.075	2.583	3.108
DI	Maximum	0.027	0.023	0.019	0.012
Xie and Benie	Minimum	0.550	0.719	0.939	1.283
EOE	Minimum	0.315	0.434	0.442	0.453

Table 2.14: Mixture models' clustering performance using the second real-life smart meter dataset

As shown in Figure 2.7b, the categorical feature representing the tariff for each energy consumer has an almost identical distribution across the clusters obtained using our proposed mixture model, having little to no influence on the energy consumption behaviour. Nevertheless, and as demonstrated by the **CH** score in Table 2.14, our proposed model has achieved clusters with a relatively small intra-cluster (within clusters) variance and a relatively large inter-cluster (between clusters) variance. Additionally, the minimum number of members within the clusters achieved using the

FSBAGGMM is 225 energy consumers, as demonstrated in Figure 2.7a. Additionally, Table 2.15 demonstrates the average values of several features for the inferred household energy consumer clusters.



(a) Percentage of energy consumers in each cluster.

(b) The distribution of tariffs across the resulting clusters

Figure 2.7: The UK power networks smart meter data clusters information

Since the smart meter data have been modelled successfully, the proposed model is capable of identifying energy consumer clusters that are suitable for demand reduction initiatives within several utility programs [74]. As an example, Table 2.15 demonstrates that the first cluster has a relatively high evening RAP with a relatively low mean STD, seasonal score, and WD-WE difference score. The power demand of energy consumers exhibiting energy consumption patterns similar to the first cluster could be lowered by implementing storage devices. The third and fourth cluster's energy consumption patterns exhibit relatively low variability in demand, as represented by the mean STD and WD-WE difference score, while exhibiting a relatively high seasonal difference in power demand, as represented by the seasonal score. Such households could be offered non-electric or more efficient heating systems to reduce the winter demand.

Consumption Profile	**Overnight RAP**	**Breakfast RAP**	**Daytime RAP**	**Evening RAP**	**Mean STD**	**Seasonal score**	**WD-WE diff. Score**
1	0.686	0.937	1.041	1.344	0.810	0.883	0.458
2	0.664	1.050	0.956	1.411	1.127	1.025	1.557
3	0.672	0.959	1.011	1.381	0.974	2.062	0.553
4	0.860	0.981	0.916	1.249	1.169	4.445	0.591

Table 2.15: The mean values of the first seven smart meter data features

2.5 Conclusion

In this chapter, an expectation-maximization algorithm is presented within the MML criterion to optimize the parameters of the bounded asymmetric generalized Gaussian mixture model and to find the optimal number of consumption profiles and the optimal subset of features simultaneously. Our approach assumes that the data arise from a mixture of bounded asymmetric generalized Gaussian distributions. The final results demonstrated that the load curve of an individual energy consumer showed a probabilistic association with each class indicating which pattern of electricity use was more or less likely to be used within a household. Therefore, it is possible to categorize households and how they consume energy using our proposed model.

Prior works in household energy consumption segmentation unrealistically approach model selection and feature selection as independent problems. Our approach successfully achieves the discovery of the true number of energy consumption profiles and the determination of the optimal set of data attributes to be used for modelling in our proposed mixture model in a single optimization process and avoids running the EM algorithm many times.

Clustering synthetically generated smart meter records with ground-truth cluster size, our proposed algorithm has outperformed most of the existing model selection approaches. In the same experiment, the proposed model correctly models the first and the second synthetic smart meter data with high accuracy of 95.569% and 91.856%, respectively. Similarly, our algorithm has also determined the optimal number of clusters in both datasets in experiments involving real-life data, and the proposed model outperforms all the mixture models used in the comparison, as demonstrated by all the utilized performance metrics. Thus, the superiority of the proposed algorithm in modelling smart meter data with different feature extraction methods over all the state-of-the-art clustering algorithms used in the comparison is proven.

Hence, our approach to analyzing real-life smart meter data is effective in determining households that are suitable for demand reduction initiatives such as DR and EE. Thus, providing the opportunity for utility companies to adopt environmentally friendly and cost-effective technologies.

Chapter 3

Explainable Finite Mixture of Mixtures of Bounded Asymmetric Generalized Gaussian and Uniform Distributions Learning for Energy Demand Management

In the previous chapter, we explored the potential of incorporating the bounded asymmetric generalized Gaussian distribution in mixture models and implemented a feature selection framework in the context of the novel mixture model. The clustering efficacy of the proposed framework has been proven successful in modelling smart meter data. Additionally, the proposed model has been able to outperform several state-of-the-art clustering models in the same context. In this chapter, and following our introduction of the principles of the mixture of mixtures, the semi-supervised learning, and the explainability in Section 1.1, Section 1.2, and Section 1.4 respectively, the potential of incorporating a Uniform distribution within the inner mixture of the mixture models is explored to build a mixture model that is reliable, explainable and robust to outliers.

We introduce a mixture of mixtures of bounded asymmetric generalized Gaussian and uniform distributions. Based on this framework, we propose model-based classification and model-based clustering algorithms. We develop an objective function for the minimum message length (MML) model selection criterion to discover the optimal number of clusters for the unsupervised approach of our proposed model. Given the crucial attention received by Explainable AI (XAI) in recent years, we introduce a method to interpret the predictions obtained from the proposed model in both learning settings by defining their boundaries in terms of the crucial features. Integrating explainability within our proposed algorithm increases the credibility of the algorithm's predictions since it would be explainable to the user's perspective through simple If-Then statements using a small binary decision tree. In this chapter, the proposed algorithm proves its reliability and superiority to several state-of-the-art machine learning algorithms within the following real-world applications: fault detection and diagnosis (FDD) in chillers, occupancy estimation and categorization of residential energy consumers.

3.1 Introduction

Statistical modelling is needed in several areas, such as pattern recognition and machine learning. In statistical learning, discovering valuable data and patterns in data relies on selecting an appropriate model to fit the data. Once the complex patterns are modelled, the trained models can be used to make valuable decisions related to the corresponding applications. Finite mixtures offer statistical models that can be trained in a supervised, unsupervised and semi-supervised manner. Finite mixture models have been considered a reliable statistical approach that can fulfil the requirements of diverse real-life applications. Mixture modelling enables using prior knowledge to model the uncertainty about the data. The term ”uncertainty” in the context of mixture models is represented by the responsibility of each mixture component to the data instances.

The Gaussian distribution as an isotropic probability density function can be used effectively within a clustering algorithm to compactly model and represent the intrinsic grouping of the data. The representation of the Gaussian mixture model comprises a set of parameters that would not relatively yield a computationally expensive model as the dimensions of the data grow. The set of parameters of each mixture component describes each discovered pattern's properties and includes a mean parameter vector and a covariance matrix. Given this motivation, the EM algorithm has been used to efficiently estimate the parameters of a Gaussian mixture model that best fits the data in several applications [19, 20]. Although this compact representation may yield clustering algorithms that require a relatively low computational cost, in some applications, it could introduce several limitations to modelling the data, such as:

(1) It has a rigid bell shape and a tail that is too short for most real-world problems [21].

(2) The Gaussian distribution and several other choices of distributions for a given mixture model are unbounded with a support range that extends from $-\infty$ to ∞ [150].

(3) The distribution is symmetric around its mean.

The data clusters of real-life applications are usually bounded and most likely have a density that is non-Gaussian [24–26]. The term 'non-Gaussian' in the context of describing a distribution signifies a density function that is asymmetric, non-bell shaped or both. The fixed kurtosis of the Gaussian distribution makes the mixture model vulnerable to the outliers of individual data clusters. Since the Gaussian distribution does not have a parameter that controls the distribution's tails to be correctly estimated while fitting the distribution to each data cluster, the distribution is unsuitable for assigning a relatively low probability of occurrence to the individual class outliers [28,29]. Such limitation brings about applications incorporating the Gaussian mixture model to seek outlier detection and removal techniques within their workflow, which in turn causes the incurrence of an additional computational expense [30–33]. On the other hand, there have been several attempts to use more flexible distributions than the Gaussian distribution to fit data from diverse applications in addition to its ability to generalize to the Gaussian distribution. Several research papers have proposed using the generalized Gaussian distribution (GGD) in diverse real-life applications [16, 40, 147, 151–153]. The distribution is formalized similarly to Equation 9. where $A(\lambda) = \left[\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}\right]^{\lambda/2}$, and X denotes the random variable. The parameters μ and σ denote the mean and the standard deviation, respectively. The symbol λ denotes the parameter that distinguishes this distribution from the Gaussian distribution. The λ parameter obeys the condition $\lambda \geq 1$ and controls the kurtosis of the distribution at which it determines if the probability density function is peaked or flat. As the parameter λ increases in value, the distribution becomes flattered until it degrades to its special case of the

Uniform distribution as $\lambda \rightarrow \infty$. As the λ decreases in value, the probability density function becomes more peaked. As $\lambda \rightarrow 0$, the probability density function becomes a delta function with an infinite value at μ. Practically, the λ parameter controls the tails of the distribution, and if estimated correctly for the training data, the distribution should model the data accurately with robustness to outliers. The distribution generalizes to the Gaussian distribution when $\lambda = 2$ and to the Laplacian distribution when $\lambda = 1$. Thus, the distribution is flexible and able to fit diverse real-life application data better in comparison to the use of each of its special cases individually in a mixture model. This distribution is especially effective if the data features are assumed to be distributed independently.

Furthermore, to fit asymmetrically distributed data, researchers have proposed the usage of the asymmetric generalized Gaussian distribution (AGGD) in several applications [123, 154, 155]. The AGGD is formalized similarly to Equation 10. In comparison to the GGD, this distribution has two parameters to describe the variance of the data instead of one parameter. The left and right standard deviations are denoted as σ_l and σ_r, respectively; they control the asymmetry of the probability density function. The distribution is skewed left if $\sigma_l < \sigma_r$, skewed right if $\sigma_l > \sigma_r$ and symmetric if $\sigma_l = \sigma_r$. Thus, in addition to the flexibility provided by the GGD, the AGGD is capable of modelling data with an asymmetrical distribution [156].

For every model-based clustering algorithm, and since the training is done in an unsupervised manner, the number of clusters is unknown. Thus, accurately modelling the data requires the discovery of the true number of classes within it. From a computational point of view, methods performing this task are categorized into deterministic and stochastic methods. Markov chain Monte Carlo (MCMC) is an example of a stochastic model selection method. It is an accurate and effective method to fit mixture models by sampling all the possible values of the mixture parameters from the full a posteriori distribution [157]. Consequently, this method is computationally expensive. On the other hand, deterministic methods attempt to discover the true number of clusters by running a candidate model for a cluster count within the range (2, M), where M represents the maximum number of clusters considered for the learning task. Akaike's information criterion (AIC) [158], Bayesian information criterion (BIC), and minimum description length (MDL) are examples of deterministic methods. The minimum message length (MML) criterion [139] is well known to outperform the BIC and the AIC model selection criteria [134–136]. Thus, we use it to discover the true number of clusters within our unsupervised learning approach of the proposed mixture model. Within the previously mentioned model selection criterion, we use the expectation-maximization (EM) to estimate the proposed model's parameters that maximize the data likelihood.

3.1.1 Explainable Artificial Intelligence

Providing explicit representations of the detected patterns by machine learning algorithms has been recently a central research topic [67–69]. There have been several research advancements in the interpretability of data-driven models. However, most of the prior works within this field have been done within post-modelling explainability. This approach has the following shortcomings [70]:

(1) It had provided no insights into training data.

(2) It was heavily dependent on the given model.

Within the publications of the last decade, the use of the term ”explainability” grew exponentially starting three years ago [71]. Our proposed model provides insight in terms of recognizing the different patterns within the data and the statistical properties of each pattern. Moreover, the integrated explainability within our proposed model further provides insights into the model and its prediction by defining the boundaries between the discovered patterns in terms of the important data attributes. Similar to the approach in [70], we develop a pre-modelling explainability in the context of model-based classification and clustering. We use this approach to justify and demonstrate in human language why specific observations are predicted within the same or different discovered patterns; This is done by training a small binary threshold tree. The adopted decision tree (DT) has a number of leaves that is equal to the number of clusters assumed within our proposed mixture model. As explained before, mixture models attempt to learn the underlying distribution, and that helps the approach to generalize well to unseen data, unlike supervised models such as DT. As well known, the predictions of the DT algorithm are easily interpretable using simple If-Then rules. Thus, integrating these two models helps us propose an explainable prediction model that generalizes well to unseen data. We further utilize the integration of explainability with our proposed model by training the mixture model using a low-level set of attributes and interpreting its predictions using a different and a high-level set of attributes that an expert of the system better understands. Prior works in explainable methods usually seek a trade-off between prediction accuracy and interpretability by varying the values of some DT hyperparameters. However, expanding the DT in some cases does not change the If-Then rules for some clusters since it yields no improvements in the cost gain [72]. Additionally, adjusting the DT hyperparameters may increase its accuracy but, at the same time, makes it less interpretable. Thus, to present the explainability of our model, we limit the scope of the experiments done within this chapter by setting the number of integrated DT leaves equal to the true number of clusters. In addition to the validation of our proposed mixture model, we demonstrate the integrated explainability through tree figures of the If-Then rules of the important features that lead to every categorization. The applications of chiller fault detection and diagnosis, energy consumers' categorization and occupancy estimation are used within this chapter to demonstrate the integrated explainability within our proposed framework. Fault diagnosis in chillers resembles the interpretation of fault categorization in terms of the attributes used. Five faults were successfully identified via a rule-based statistical model for vapour compression air conditioners [159]. In the previously mentioned research paper, the authors have provided interpretability of what characterizes the faults in terms of the data attributes. The fault categories were explained through a simultaneous increasing or decreasing change to seven different attributes instead of specifying thresholds. Thus, in addition to our aim to classify faults in chiller operational data, we aim to provide interpretability in the form of If-Then statements using the values of the data features.

Our approach to analyzing households' energy consumption data determines households suitable for demand reduction initiatives such as demand response (DR) and energy efficiency (EE). Thus, providing the opportunity for utility companies to adopt environmentally friendly and cost-effective technologies. DR is an incentive program that enables the possibility for utility companies to reduce expenses on unnecessary investments and lower emissions of greenhouse gases (GHG). DR induces households to reduce their energy consumption levels at times of high wholesale market prices or when system reliability is jeopardized. EE programs aim to reduce the power demand of households while maintaining their consumption habits. As an example, the power demand of energy consumers exhibiting a relatively high evening demand with low variability can lower their peak load by implementing storage devices. As households exhibiting relatively low variability in demand with a relatively high seasonal difference in power demand, they could be offered non-electric or more efficient heating systems to reduce the winter demand. Prior works to target energy consumers using clustering algorithms [73, 74] use additional statistical analysis to select suitable candidate energy consumers for demand management programs [75–77]. Thus, we take advantage of the integration of the explainability within our proposed method and define the boundaries between energy consumers using simple If-Then statements with specific values of high-level features so it would be easy for an expert to find a suitable candidate for any power demand reduction program. Besides the successful efforts in correctly detecting the patterns and their corresponding number of occupants, researchers have exhibited the estimated occupancy against the true occupancy with respect to a single dimension or a max of two dimensions [160, 161]. However, in this chapter, we aim to present the estimated occupancy of our proposed model with respect to the values of important features with a simple graph of If-Then rules.

In this chapter, we propose a mixture of mixtures of bounded asymmetric generalized Gaussian and Uniform distributions (BAGGU) that can generalize to an extensive range of mixture models. The inner mixture of the proposed framework consists of a flexible distribution that can generalize to several distributions and a Uniform distribution that can help increase the model's robustness to outliers. We propose semi-supervised and unsupervised methods of learning the proposed model. Using the unsupervised approach, we prove that our proposed model is capable of performing pattern recognition by inferring the true number of energy consumption profiles and identifying household groups that follow each discovered pattern using the **Minimum Message Length** (MML) model selection criterion. Using several performance metrics, our mixture model outperforms several state-of-the-art machine learning models such as the Explainable **mixture of mixtures of Gaussian and Uniform distribution** (ExGU), and the Explainable **ExKMC** in modelling households' energy consumption data. For the semi-supervised learning approach, we validate our proposed model against two interesting real-life applications: Chiller fault detection and diagnosis and Occupancy estimation.

Chiller fault detection is an AI-driven application for identifying and diagnosing issues in large-scale cooling systems, optimizing performance and reducing energy consumption. By utilizing machine learning and data analytics, AI models can analyze historical and real-time data from chiller components to detect anomalies and predict potential faults. Occupancy estimation is an AI-based application that predicts the number of people present in a given space, such as buildings or rooms, by analyzing data from various sensors and sources [162]. Leveraging machine learning algorithms, this approach aids in optimizing energy usage, enhancing security, and improving building management operations.

Using several performance metrics, our mixture model outperforms the Explainable **mixture of mixtures of Gaussian and Uniform distribution** (ExGU), the Explainable Adaptive Boosting (ExAdaboost) algorithm with DT as its base estimator, the Explainable **k-Nearest Neighbour** (ExKNN), and the Decision Tree within both real-life applications. Within the subsequent sections of this chapter, the information is presented with the following arrangement: Section 3.2 describes the proposed mixture model, its semi-supervised learning approach and its unsupervised learning approach. Section 3.3 introduces the real-life applications used to validate the proposed model, their datasets and their experimental results. Section 3.4 presents our conclusions and future works.

3.2 The proposed mixture model

3.2.1 The finite mixture model

Mixture models offer a powerful clustering solution for diverse applications. The basic concept of mixture models is that they assume the data arise from a convex combination of distributions, where each cluster g is represented by a single distribution $f(\vec{X}_i|\xi_g)$ with parameters ξ_g within the mixture. A parametric M-component mixture model is defined similarly to Equation 1. where $\vec{X} = [X_d, ..., X_d]^T$ denotes a multivariate random vector. The term π_g represents the mixing proportion of the component g and it fulfils the following conditions: $\pi_g > 0$, $\sum_{g=1}^{M} \pi_g = 1$. $\Theta = (\pi_1, ..., \pi_M, \xi_1, ..., \xi_M)$ denotes the complete parameter set that defines the mixture model.
In this chapter, we are proposing a mixture of mixtures where each component of the mixture model

is itself a mixture of two distributions. We considered the inner mixture to consist of a bounded asymmetric generalized Gaussian $\phi(\vec{X}|\vartheta_g)$ and a Uniform distribution $u(\vec{X}|\zeta_g)$. Consequently, each component of the mixture is defined as follows:

$$f(\vec{X}_i|\xi_g) = \omega_g \phi(\vec{X}_i|\vartheta_g) + (1-\omega_g)u(\vec{X}_i|\zeta_g) \tag{51}$$

The bounded asymmetric generalized Gaussian distribution is defined following [22, 102] in a similar manner to Equation 8:

where $\Psi(X_{id}|\theta_{gd})$ denotes the unbounded asymmetric generalized Gaussian probability density function.

The terms $H(X_{id}|g)$ and $\int_{\partial_k} \Psi(X|\theta_{gd})dX$ contribute to making the AGGD used in the proposed inner mixture bounded. The bounded support region is denoted by τ_{gd} for each component g and dimension d. The indicator function $H(X_{id}|g)$ sets the density value of the AGGD outside the bounded support region to zero, and it is defined similarly to Equation 7. The term $\int_{\tau_{gd}} \Psi(X|\theta_{gd})dX$ denotes the normalization constant that restores the statistical properties of the probability density function, the share of $\Psi(X_{id}|\theta_{gd})$ that falls within the support region τ_{gd}. In comparison to the Gaussian distribution, the additional parameters of the BAGGD allowed the proposed mixture model to be flexible and able to fit data with different shapes, asymmetry and bounded support. Additionally, the Uniform distribution within the inner mixture makes the proposed mixture model more robust to outliers.

3.2.2 Semi-supervised learning of the mixture parameters

In this section, we will explain how we train the model in a semi-supervised manner. The data's likelihood considering N D-dimensional observations where s observations are labelled, or their cluster memberships are assumed known is defined as follows:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{s}\prod_{g=1}^{M}\left[\pi_g f(\vec{X}_i|\xi_g)\right]^{Z_{ig}} \times \prod_{j=s+1}^{N}\sum_{h=1}^{M}\pi_h f(\vec{X}_j|\xi_g) \tag{52}$$

where Z_i is of standard basis and it indicates which cluster k is mostly responsible of observation

i. Thus, Z_i is a set of all possible latent variables $Z_{i1}, ..., Z_{iM}$ for each observation i. In order to find the optimal parameters that maximize the data likelihood as computed in Equation 52, we use the EM algorithm to achieve this task in an iterative manner. The objective function of the EM algorithm in each iteration considering the semi-supervised approach of learning the proposed mixture of mixtures is written as follows:

$$\begin{aligned}
\mathcal{L}(\mathcal{X},&\Theta, Z, V) = \\
&\sum_{i=1}^{s}\sum_{g=1}^{M} Z_{ig}\big[\log \pi_g + \hat{V}_{ig}\log\phi(\vec{X}_i|\vartheta_g) \\
&+ (1-\hat{V}_{ig})\log u(\vec{X}_i|\zeta_g)\big] \\
&+ \sum_{j=k+1}^{N}\sum_{h=1}^{M}\gamma(Z_{jh})\big[\log \pi_h \\
&+ \hat{V}_{jh}\log\phi(\vec{X}_j|\vartheta_h) + (1-\hat{V}_{jh})\log u(\vec{X}_j|\zeta_g)\big]
\end{aligned} \tag{53}$$

The conditional expected values are denoted as $\hat{Z}_{jh}$ and $\hat{v}_{ig}$. $\hat{Z}_{jh}$ is defined in Equation 54 for $j = k+1, ..., N$. $\hat{v}_{ig}$ is defined in Equation 55 for $i = 1, ..., N$.

$$p(Z_j = g|\vec{X}_j, \Theta) = \gamma(Z_{jh}) = \frac{\hat{\pi}_h f(\vec{X}_j|\hat{\xi}_h)}{\sum_{g=1}^{M}\hat{\pi}_g f(\vec{X}_j|\hat{\xi}_g)} \tag{54}$$

$$\hat{V}_{ig} = \frac{\hat{\omega}_g\phi(\vec{X}|\vartheta_g)}{\hat{\omega}_g\phi(\vec{X}|\vartheta_g) + (1-\hat{\omega}_g)\phi(\vec{X}|\vartheta_g)} \tag{55}$$

In order to optimize the data's likelihood, we differentiate Equation 53 with respect to each one of the model parameters and set it to zero to obtain the following Equations:

$$\pi_g = p(Z_g = 1) = \frac{\sum_{i=1}^{s} Z_{ig} + \sum_{j=k+1}^{N}\hat{Z}_{jg}}{N} \tag{56}$$

$$\omega_g = \frac{\sum_{i=0}^{s}\hat{V}_{ig}Z_{ig} + \sum_{j=s+1}^{N}\hat{V}_{jg}\hat{Z}_{jg}}{\sum_{i=1}^{s} Z_{ig} + \sum_{j=s+1}^{N}\hat{Z}_{jg}} \tag{57}$$

As we calculate the gradient of Equation 53 with respect to each parameter of our proposed model, we obtain non-linear equations. Thus, we use the Newton–Raphson method to approximate the updated values of the model parameters for each iteration of the EM algorithm. The model's parameters are estimated as follows:

$$\hat{\mu}_{gd} = \mu_{gd} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \mu_{gd}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \mu_{gd}} \right) \right] \tag{58}$$

$$\hat{\sigma}_{l_{gd}} = \sigma_{l_{gd}} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{l_{gd}}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{l_{gd}}} \right) \right] \tag{59}$$

$$\hat{\sigma}_{r_{gd}} = \sigma_{r_{gd}} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{r_{gd}}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{r_{gd}}} \right) \right] \tag{60}$$

$$\hat{\lambda}_{gd} = \lambda_{gd} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \lambda_{gd}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \Theta, Z, V)}{\partial \lambda_{gd}} \right) \right] \tag{61}$$

The partial derivatives are developed in Appendix A.2.

3.2.3 Unsupervised maximum likelihood estimation of the mixture parameters

The unsupervised learning approach of a mixture model is considered a special case of the semi-supervised learning approach. Thus, Setting $(s = 0)$ in Equation 52 yields the data likelihood of the unsupervised clustering approach as follows:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{h=1}^{M} \pi_h f(\vec{X}_i|\xi_h) \tag{62}$$

As a normal procedure, we would be able to compute the gradient of Equation 62 with respect to the model parameters $\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \theta}$ and set it to 0, then we solve for Θ. However, solving for Θ would not yield a closed-form solution. Thus, we define the complete data log-likelihood as follows:

$$p(\mathcal{X}, Z, V|\Theta) = \prod_{j=1}^{N} \prod_{h=1}^{M} \left[\pi_h \phi(\vec{X}_j|\vartheta_h)^{V_{jh}} u(\vec{X}_j|\zeta_g)^{(1-\hat{V}_{jh})} \right]^{Z_{jh}} \tag{63}$$

The complete data likelihood is used to form the objective function of the EM algorithm at each iteration as follows:

$$\mathcal{Q}(\mathcal{X},\Theta,Z,V)=\sum_{j=1}^{N}\sum_{h=1}^{M}\gamma(Z_{jh})\log(\pi_h f(\vec{X}_j|\xi_h)) \tag{64}$$

$$\begin{aligned}\mathcal{Q}(\mathcal{X},\Theta,Z,V)=&\sum_{j=1}^{N}\sum_{h=1}^{M}\gamma(Z_{jh})\big[\log\pi_h+\hat{V}_{jh}\log\phi(\vec{X}_j|\vartheta_h)\\&+(1-\hat{V}_{jh})\log u(\vec{X}_j|\zeta_g)\big]\end{aligned} \tag{65}$$

The conditional expected values $\gamma(Z_{jh})$ and $\hat{V}_{jh}$ are given by Equations 66 and 67.

$$p(Z_j=h|\vec{X}_j,\Theta)=\gamma(Z_{jh})=\frac{\hat{\pi}_h f(\vec{X}_j|\hat{\xi}_h)}{\sum_{g=1}^{M}\hat{\pi}_g f(\vec{X}_j|\hat{\xi}_g)} \tag{66}$$

$$\hat{V}_{jh}=\frac{\hat{\omega}_h\phi(\vec{X}|\vartheta_h)}{\hat{\omega}_h\phi(\vec{X}|\vartheta_h)+(1-\hat{\omega}_h)\phi(\vec{X}|\vartheta_h)} \tag{67}$$

In a restricted maximization process, we find the updated values of the outer mixing proportion π_g and the inner mixing proportions ω_g for each cluster g as follows:

$$\pi_h=p(Z_h=1)=\frac{\sum_{i=1}^{N}p(Z_h=1|\vec{X}_i,\Theta_M)}{N} \tag{68}$$

$$\omega_h=\frac{\sum_{j=1}^{N}\hat{V}_{jh}\hat{Z}_{jh}}{\sum_{j=1}^{N}\hat{Z}_{jg}} \tag{69}$$

Furthermore, we evaluate the first and second partial derivatives of Equation 65 to find the updated values of the mixture model's parameters at the M-step of each iteration of the EM algorithm as follows:

$$\hat{\mu}_{hd}=\mu_{hd}-\left[\left(\frac{\partial^2\mathcal{Q}(\mathcal{X},\Theta,Z,V)}{\partial\mu_{hd}^2}\right)^{-1}\left(\frac{\partial\mathcal{Q}(\mathcal{X},\Theta,Z,V)}{\partial\mu_{hd}}\right)\right] \tag{70}$$

$$\hat{\sigma}_{l_{hd}}=\sigma_{l_{hd}}-\left[\left(\frac{\partial^2\mathcal{Q}(\mathcal{X},\Theta,Z,V)}{\partial\sigma_{l_{hd}}^2}\right)^{-1}\left(\frac{\partial\mathcal{Q}(\mathcal{X},\Theta,Z,V)}{\partial\sigma_{l_{hd}}}\right)\right] \tag{71}$$

$$\hat{\sigma}_{r_{hd}} = \sigma_{r_{hd}} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma^2_{r_{hd}}} \right)^{-1} \left(\frac{\partial \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{r_{hd}}} \right) \right] \tag{72}$$

$$\hat{\lambda}_{hd} = \lambda_{hd} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \lambda^2_{hd}} \right)^{-1} \left(\frac{\partial \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \lambda_{hd}} \right) \right] \tag{73}$$

3.2.4 Model selection for unsupervised learning

Since this is an unsupervised learning approach, the number of components is assumed to be unknown even if labels are available. The MML approach of finding the true number of mixture components is based on evaluating a set of statistical models. Each model is evaluated based on its ability to compress a message containing the data. The message used to evaluate each model in the set of candidate models consists of two parts. The first part of the message encodes information exclusively about the candidate model, and it consists of prior information about its parameters. The second part encodes information about the training data exclusively in an approach that takes the model into account. In order to implement MML, the candidate model that minimizes the following objective function specifies the optimal number of mixing components:

$$\begin{aligned} \text{messLen} \approx & - \log(P(\Theta_M)) - \mathcal{Q}(\mathcal{X}, \Theta_M, Z, V) + \frac{1}{2} \log |F(\Theta)| \\ & + \frac{\eta}{2}(1 + \log(h_\eta)) \end{aligned} \tag{74}$$

where $p(\Theta_M)$ is the prior probability that expresses the lack of knowledge about the mixture parameters with M components, $F(\Theta_M)$ is the Fisher information matrix, η is the number of the model's free parameters, and h_η is the optimal quantization lattice constant $\mathbb{R}^\eta$ [141]. The number of free parameters (η) used to define our proposed mixture model can be calculated as follows:

$$\eta = M(5D + 1) \tag{75}$$

Prior probability $P(\Theta_M)$

As mentioned earlier, this prior probability expresses the lack of knowledge about the proposed mixture model's parameters. Since the knowledge about the parameters of a specific mixture component does not provide any information about the parameters of the other different components, it is safe to assume that the parameters of the different mixture components are independent. Consequently, the prior probability is calculated as follows:

$$P(\Theta_M) = P(\mu)P(\lambda)P(\sigma_l)P(\sigma_r)P(\pi) \tag{76}$$

where the mixture parameters ($\mu = \{\mu_g\}, \sigma_l = \{\sigma_{l_g}\}, \sigma_r = \{\sigma_{r_g}\}, \lambda = \{\lambda_g\}, \pi = (\pi_1, ..., \pi_M)$) are considered mutually independent. Since the parameter π obeys the condition $\sum_{g=1}^{M} \pi_g = 1$, we use the Dirichlet distribution to compute its prior as follows:

$$P(\pi) = \frac{\Gamma\big(\sum_{g=1}^{M} \kappa_g\big)}{\prod_{g=1}^{M} \Gamma(\kappa_g)} \prod_{g=1}^{M} \pi_g^{(\kappa_g - 1)} \tag{77}$$

where Γ denotes the Gamma function and $(\kappa_1, \kappa_2, ..., \kappa_M)$ denotes the parameters of the Dirichlet distribution. Assuming a uniform prior where $\kappa_1, \kappa_2, ...\kappa_M = \kappa = 1$, the prior of the parameter π is evaluated as follows:

$$P(\pi) = (M-1)! \tag{78}$$

For each μ_{gd}, we're assuming a uniform prior within the region ($\mu_d - \sigma_{l_d} \leq \mu_{gd} \leq \mu_d + \sigma_{r_d}$), where $\vec{\mu} = (\mu_1, \ldots, \mu_D)$, $\vec{\sigma_l} = (\sigma_{l_1}, \ldots, \sigma_{l_D})$, and $\vec{\sigma_r} = (\sigma_{r_1}, \ldots, \sigma_{r_D})$ denote the mean, left standard deviation, and right standard deviation vectors of the sample data. Consequently, the prior of μ is calculated as follows:

$$P(\mu) = \prod_{g=1}^{M} \prod_{d=1}^{D} P(\mu_{gd}) = \prod_{d=1}^{D} \frac{1}{(\sigma_{l_d} + \sigma_{r_d})^M} \tag{79}$$

As for the parameter λ, we assume that its prior can be described using a uniform distribution within the range $[0 - \tau]$, where τ is the largest value permitted for the parameter. Thus, the prior for λ is written as follows:

$$p(\lambda) = \prod_{g=1}^{M} \prod_{d=1}^{D} P(\lambda_{gd}) = \frac{1}{\gamma^{MD}} \tag{80}$$

Similarly, a uniform distribution is considered for both parameters σ_l and σ_r considering the following ranges: $0 \leq \sigma_{l_{gd}} \leq \sigma_{l_g}$, $0 \leq \sigma_{r_{gd}} \leq \sigma_{r_g}$. Thus the prior for the parameters σ_l and σ_r are given as follows:

$$P(\sigma_l) = \prod_{g=1}^{M}\prod_{d=1}^{D} P(\sigma_{l_{gd}}) = \prod_{d=1}^{D} \frac{1}{\sigma_{l_d}^{M}} \tag{81}$$

$$P(\sigma_r) = \prod_{g=1}^{M}\prod_{d=1}^{D} P(\sigma_{r_{gd}}) = \prod_{d=1}^{D} \frac{1}{\sigma_{r_d}^{M}} \tag{82}$$

Estimating the Fisher information matrix

The Fisher information matrix is the expected value of the Hessian matrix given the negative log-likelihood. It is difficult to analytically evaluate this expected value given our proposed mixture model. Thus, the Hessian matrix can be approximated via the calculation of the complete Fisher information matrix as shown in Equation 83.

$$|F(\Theta)| = |F(\pi)| \prod_{g=1}^{M} |F(\vec{\mu}_g)||F(\vec{\sigma}_{l_g})||F(\vec{\sigma}_{r_g})||F(\vec{\lambda}_g)| \tag{83}$$

where $|F(\pi)|, |F(\vec{\mu}_g)|, |F(\vec{\sigma}_{l_g})|, |F(\vec{\sigma}_{r_g})|,$ and $|F(\vec{\lambda}_g)|$ represent the determinants of the Fisher information matrices with respect to the parameter vectors $\vec{\mu}_g, \vec{\sigma}_{l_g}, \vec{\sigma}_{r_g}$ and $\vec{\lambda}_g$ respectively. $|F(\pi)|$ is computed as follows:

$$|F(\pi)| = \frac{N^{(M-1)}}{\prod_{g=1}^{M} \pi_g} \tag{84}$$

In consideration of the parameter vectors $\vec{\mu}_g, \vec{\sigma}_{l_g}, \vec{\sigma}_{r_g}, \vec{\lambda}_g$, the Hessian matrices are computed as follows:

$$F(\vec{\mu}_g)_{k1,k2} = \frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \mu_{gd_1} \partial \mu_{gd_2}} \tag{85}$$

$$F(\vec{\sigma}_{l_g})_{k1,k2} = \frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{l_{gd_1}} \partial \sigma_{l_{gd_2}}} \tag{86}$$

$$F(\vec{\sigma}_{r_g})_{k1,k2} = \frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \sigma_{r_{gd_1}} \partial \sigma_{r_{gd_2}}} \tag{87}$$

$$F(\vec{\lambda}_g)_{k1,k2} = \frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Theta, Z, V)}{\partial \lambda_{gd_1} \partial \lambda_{gd_2}} \tag{88}$$

where $(d_1, d_2) \in (1, ..., D)$. The partial derivatives listed in Appendix A.2 can be used to evaluate Equations (85)-(88).

3.2.5 Convergence of the EM algorithm

At each iteration of the EM algorithm within both the unsupervised and semi-supervised learning settings, we adopt the Aitken acceleration [163] to estimate the asymptotic maximum of the log-likelihood. Based on the recently mentioned estimated value, we establish a condition for the convergence of the EM algorithm. The estimated parameters within the converging iteration are the optimal parameters that fit the proposed model to the training data. For iteration ι, the Aitken acceleration is estimated as follows:

$$a = \frac{p(\mathcal{X}|\Theta)^{\iota+1} - p(\mathcal{X}|\Theta)^{\iota}}{p(\mathcal{X}|\Theta)^{\iota} - p(\mathcal{X}|\Theta)^{\iota-1}} \tag{89}$$

The asymptotic estimate of the log-likelihood at iteration $\iota + 1$ is calculated at follows:

$$p(\mathcal{X}|\Theta)_{\infty}^{\iota+1} = p(\mathcal{X}|\Theta)^{\iota} + \frac{p(\mathcal{X}|\Theta)^{\iota+1} - p(\mathcal{X}|\Theta)^{\iota}}{1 - a^{(\iota)}} \tag{90}$$

3.2.6 Integration of explainability

In our proposed framework, we leverage the power of decision trees to provide an explainable machine-learning model. This section aims to elucidate the process of integrating decision trees within our framework, which allows us to obtain the If-Then rules that enable us to better understand and explain the model's predictions.

(1) Training Decision Trees using Discovered Patterns

The foundation of our framework is built on the patterns discovered by our proposed mixture model. These patterns serve as the basis for training a decision tree, which in turn, learns to generate If-Then rules based on these identified patterns. This enables us to obtain an interpretable and explainable model, providing a clear understanding of the underlying logic behind the predictions.

(2) Decision Tree Predicted Labels

Once the decision tree is trained using the patterns discovered by our mixture model, it can be employed to make predictions on new data. These predicted labels are derived from the

If-Then rules generated during the decision tree's training process.

(3) Decision Tree Validation

To assess the efficacy of our proposed framework, we employ a set of performance metrics that serve as validation criteria for our proposed framework's predicted labels. By evaluating these metrics, we can gauge the success of our framework in generating accurate and interpretable predictions.

In summary, the integration of decision trees within our proposed framework enables us to create an explainable machine-learning model, providing valuable insights into the decision-making process. By training the decision tree using patterns discovered by our mixture model, we can obtain If-Then rules that offer a transparent understanding of the model's predictions. Furthermore, we validate the application of our framework through the use of performance metrics, ensuring its effectiveness in generating accurate and interpretable results.

3.3 Experimental results

In this section, we demonstrate the effectiveness of our proposed model using three energy management-related applications, namely chiller fault detection and diagnosis, occupancy estimation, and energy consumption categorization. The performance of our method is also compared against state-of-the-art methods. Several performance evaluation metrics were introduced in [131] and well defined within Section 2.4 were used within the experiments of this chapter, and they are listed as follows:

- **S** [144]: The Silhouette score
- **CH** [145]: The **Calinski-Harabasz**
- **DB** [146]: The **Davies–Bouldin**
- **MCC**: it represents the **Mathiews Correlation Coefficient** evaluation metric [148].

Our method's performance is compared against several state-of-the-art machine learning algorithms, namely Explainable mixture of mixtures of Gaussian and uniform distributions (ExGu), Explainable Adaboost (ExAdaboost), Explainable KNN (ExKNN), and Decision Tree (DT) Within experiments of applications with labelled data. Within experiments of applications with unlabelled data, we compare the performance of our proposed model to ExGU and Explainable K-means (ExKMC).

3.3.1 Chiller fault detection and diagnosis

The energy consumption of the building sector comprises 36% of the global consumption, and it is set to increase gradually for each year in the upcoming two decades [164, 165]. Since HVAC systems comprise half the energy consumption of commercial buildings [166–168], variations in weather conditions across time and locations are considered to be a major factor that impacts energy consumption. In addition to the influence of weather, the following factors constitute the unnatural or unexpected drivers of the poor efficiency of HVAC systems: wrongful installation, components malfunction, control mistakes and unprofessional maintenance. Thus, chiller faults can cause both discomfort and inefficient power demand in the containing building and modelling chiller operation data is a necessary task. In some states, the deployment of a fault detection and diagnosis (FDD) system accompanying specific types of cooling units is mandatory. According to an overview of FDD techniques and practices [166], FDD methods are classified into three categories according to their driving bases, and they are listed as follows:

(1) A priori knowledge-based

(2) Gray box

(3) Data-driven

The effective mechanism followed within all these approaches is based on the establishment of a model and the usage of measurement data to detect faults. However, as mentioned before, they are categorized based on the difference between the methods' driving base and resources.

The FDD methods of the first category handle its task by running a comparison between the operating state obtained from the measurements and the expected operating states; the expected operating states are recorded from a model that is built using a priori knowledge (e.g., first principles) [166]. However, the boundaries between the rule-based and physical models can be blurred for some approaches under this category. Methods under this category have been widely used for the purpose of detecting and diagnosing faults in HVAC systems [169, 170]. Although their performance expectations could extrapolate very well in the case of the scarcity of the training data, it is extremely difficult and expensive to build the physical models that generate the expected operation states of the system. Thus, those methods are exclusively applicable to information-rich systems where abundant types of sensors are available to accurately register the system's behaviour in response to external physical factors. Additionally, both types of models used in this category of methods (physical models and rule-based models) are usually effective with a small number of inputs, outputs and states. As an advantage of rule-based approaches, they have been successful in providing fault diagnosis. Gray box methods are hybrids of data-driven and a priori knowledge-based methods. Gray box methods use a priori knowledge to mathematically formalize the model, and the model parameters are estimated using the measurement data. Thus, models of gray-box methods require thorough user expertise in the system and in statistics. HVAC systems of a commercial scale is a non-linear system that is challenging and time-consuming to model (using gray box and a priori knowledge-based methods) [171]. Meanwhile, models within data-driven methods rely exclusively on historical measurements obtained from the chiller to learn a mathematical model that predicts faults [172]. They are also known as data-driven methods, and they mathematically relate the historical input measurements to their expected output with no dependence on theoretical system models [172, 173]. Thus,

Performance Measure	**ExBAGGU**	**ExGU**	**ExAdaboost**	**ExKNN**	**DT**
Accuracy	98.850	96.215	96.701	96.218	95.305
Silhouette	0.040	0.015	-0.001	0.015	0.018
Calinski Harabasz	344.442	375.674	426.977	375.788	373.127
Davies Bouldin	4.699	4.750	5.268	5.741	5.313
Fowlkes Mallows	0.913	0.761	0.780	0.761	0.720
Mathiews Correlation Coefficient	94.780	82.865	85.149	82.874	79.021
F1-Score	95.400	84.861	86.802	84.871	81.220
Jaccard	0.954	0.849	0.868	0.849	0.812
ROC AUC	0.974	0.913	0.925	0.914	0.893
V Measure	0.902	0.767	0.789	0.767	0.744
Rand	0.978	0.939	0.944	0.939	0.927
Normalized Mutual Information	0.902	0.767	0.789	0.767	0.744
Mutual Info	1.874	1.591	1.633	1.590	1.528
Homogeneity	0.901	0.765	0.786	0.765	0.735
Completeness	0.902	0.770	0.793	0.770	0.753
Adjusted Rand	0.900	0.727	0.747	0.727	0.677
Adjusted Mutual Info	0.902	0.767	0.789	0.767	0.744

Table 3.1: Performance evaluation of the models used in the comparison for the detection of faults of severity level 4

those models do not require the difficult task of understanding and thus modelling the system operation. Therefore, data-driven methods have presented a robust solution to identify chiller faults when a priori knowledge-based methods fail and exhibit an outstanding potential in modelling chiller's operational data [174]. Data-driven models are trained in different manners, namely: supervised, unsupervised and semi-supervised. Supervised methods such as deep neural networks require measurement data of massive size to perform well since HVAC systems operate under widely diverse circumstances over different locations and weather conditions for each location over the year [175]. Thus, supervised models are not known to extrapolate well [166]. The underlying distribution of the chiller operational data can be discovered using mixture models. Overfitting is avoidable while learning a mixture model in an unsupervised or semi-supervised manner [163]. Bayes's rule can be used to distinguish between the chiller's operational states after the model is fit to the training data. c [165]. Explaining the reason models within data-driven methods classify different observations with the same or different labels is extremely difficult. However, this function is very useful as it resembles the function of diagnosis in the FDD application. Thus, and as mentioned in the previous sections, we aim to integrate model-based classification in order to provide a diagnosis within the FDD application used to validate our proposed model.

Performance Measure	ExBAGGU	ExGU	ExAdaboost	ExKNN	DT
Accuracy	98.668	97.428	97.839	96.721	94.207
Silhouette	0.074	0.048	0.048	0.038	0.048
Calinski Harabasz	371.937	401.374	405.726	381.919	452.419
Davies Bouldin	3.549	4.034	4.472	4.990	4.100
Fowlkes Mallows	0.940	0.845	0.856	0.794	0.729
Mathiews Correlation Coefficient	99.431	88.472	90.161	85.778	74.716
F1-Score	96.673	89.712	91.358	86.885	76.829
Jaccard	0.967	0.897	0.914	0.869	0.768
ROC AUC	0.987	0.941	0.951	0.925	0.868
V Measure	0.954	0.878	0.870	0.838	0.801
Rand	0.980	0.960	0.964	0.945	0.924
Normalized Mutual Information	0.954	0.878	0.870	0.838	0.801
Mutual Info	2.173	1.817	1.807	1.716	1.610
Homogeneity	0.952	0.874	0.869	0.825	0.774
Completeness	0.955	0.881	0.870	0.852	0.830
Adjusted Rand	0.957	0.822	0.835	0.761	0.681
Adjusted Mutual Info	0.954	0.877	0.870	0.838	0.801

Table 3.2: Performance evaluation of the models used in the comparison for the detection of faults of severity level 3

In order to validate our proposed mixture model using this application, we use the ASHRAE RP-1043 Dataset [176, 177]. The dataset consists of several observations for chiller operational data; It contains various common chiller faults. The chiller operational data was recorded at 10-second and 1-minute intervals. Our proposed mixture model was validated using the dataset with a 10-second interval. The 90-ton centrifugal water-cooled chiller was utilized to record the normal and fault data within this dataset. The previously mentioned chiller can be considered as the representative of chillers that are considered in larger installations [176]. Four datasets were used to validate the proposed model. Each dataset contains the chiller normal and faulty operational data with a different severity level. The severity levels are within the range [1-4]. For each severity level, seven types of faults are introduced, and they are listed as follows:

(1) RefLeak: Resembles a refrigerant leak and is labelled with the class number '6'

(2) RefOVer: Resembles a refrigerant overcharge and is labelled with the class number '7'

(3) ExcsOil: Resembles an excess Oil and is labelled with the class number '2'

(4) ReduCF: Resembles a reduced Condenser Water Flow and is labelled with the class number '3'

(5) ReduEF: Resembles a reduced evaporator Water Flow and is labelled with the class number '4'

(6) ConFoul: Resembles a condenser fouling and is labelled with the class number '1'

(7) NonCon: Indicates to a non-condensable in refrigerant and is labelled with the class number '5'

Within each observation of the used datasets, 65 features were recorded. The features include the following: Condenser valve position, oil feed temperature, a flow rate of condenser water, etc., which are recorded every ten seconds, as mentioned earlier. Each dataset consists of 41,528 observations, where each fault is represented by 5191 observations. Using our proposed model, we've been able to discover and learn the normal and the fault patterns that are registered within the four datasets mentioned earlier. First, the dataset was divided into two parts, one part was used for training, and the other was used for testing. Furthermore, a portion of 10% of the training dataset was presumed with known labels, and the remaining 90% of the dataset was presumed with unknown labels. Additionally, we have measured the proposed model's performance on the testing data using several performance metrics, as will be demonstrated later in this section. The labels obtained from

Performance Measure	ExBAGGU	ExGU	ExAdaboost	ExKNN	DT
Accuracy	96.340	94.162	93.568	93.701	92.186
Silhouette	0.006	-0.002	0.003	-0.014	0.015
Calinski Harabasz	326.590	419.597	446.730	430.699	486.755
Davies Bouldin	4.648	5.009	4.967	4.931	5.208
Fowlkes Mallows	0.780	0.652	0.651	0.644	0.629
Mathiews Correlation Coefficient	84.080	73.573	71.466	71.688	66.864
F1-Score	85.360	76.650	74.270	74.805	68.745
Jaccard	0.854	0.766	0.743	0.748	0.687
ROC AUC	0.916	0.867	0.853	0.856	0.821
V Measure	0.810	0.685	0.694	0.681	0.691
Rand	0.941	0.911	0.905	0.907	0.878
Normalized Mutual Information	0.810	0.685	0.694	0.681	0.691
Mutual Info	1.653	1.416	1.402	1.399	1.315
Homogeneity	0.795	0.681	0.674	0.673	0.632
Completeness	0.825	0.689	0.715	0.690	0.762
Adjusted Rand	0.744	0.601	0.595	0.590	0.546
Adjusted Mutual Info	0.809	0.685	0.694	0.681	0.691

Table 3.3: Performance evaluation of the models used in the comparison for the detection of faults of severity level 2

the models used in the comparison were used along with the testing data to train a decision tree that provides explainability, as mentioned earlier. Our proposed mixture model was able to recognize the existing patterns and achieve an accuracy of 98.85% Using the dataset with faults of severity level 4 as demonstrated in Table 3.1. The semi-supervised mixture of mixtures of Gaussian and Uniform distributions has been able to achieve an accuracy that is lower than that of our proposed mixture model.

Performance Measure	ExBAGGU	ExGU	ExAdaboost	ExKNN	DT
Accuracy	94.357	91.198	90.423	90.328	90.286
Silhouette	0.028	-0.007	-0.028	-0.025	-0.038
Calinski Harabasz	217.749	378.126	360.283	490.296	545.189
Davies Bouldin	4.082	4.176	4.319	4.879	4.620
Fowlkes Mallows	0.650	0.596	0.605	0.576	0.607
Mathiews Correlation Coefficient	76.040	63.936	62.750	57.957	62.429
F1-Score	77.428	64.792	61.693	61.313	61.144
Jaccard	0.774	0.648	0.617	0.613	0.611
ROC AUC	0.871	0.799	0.781	0.779	0.778
V Measure	0.954	0.711	0.739	0.673	0.743
Rand	0.896	0.844	0.811	0.863	0.806
Normalized Mutual Information	0.854	0.711	0.739	0.673	0.743
Mutual Info	1.653	1.319	1.309	1.280	1.306
Homogeneity	0.799	0.634	0.630	0.616	0.628
Completeness	0.952	0.809	0.893	0.741	0.910
Adjusted Rand	0.684	0.483	0.457	0.487	0.454
Adjusted Mutual Info	0.924	0.711	0.738	0.673	0.743

Table 3.4: Performance evaluation of the models used in the comparison for the detection of faults of severity level 1

Using the dataset with the severity level-3 faults, our proposed model has been able to outperform other models used in comparison with an accuracy score of 98.668% as demonstrated in Table 3.2. However, the adaptive Boosting (Adaboost) classifier, which was trained with a decision tree as a base estimator, has been able to outperform the semi-supervised mixture of mixtures of Gaussian and Uniform distributions in this experiment with an additional accuracy of 0.411%.

Similarly, for datasets with severity levels 1 and 2, our proposed mixture model has been able to outperform all the other machine learning models used in the comparison, as shown in Table 3.4 and Table 3.3 respectively.

As explained earlier, regardless of the proven effectiveness of our proposed mixture model in detecting patterns, it does not provide an explicit representation of the detected patterns. The integration of explainability within our proposed method using DT produces the following Figures 3.1, 3.2, 3.3, and 3.4 for datasets containing chiller faults of severity 1, 2, 3 and 4 respectively. Given Table 3.5 and the previous explanation of the pattern labels in this section, we present how the data instances are divided according to the values of the important features. As demonstrated in Figure 3.1 we can conclude the exact feature values indicating that the chiller is running with no fault. The data attributes Condenser Valve Position and Evaporator Valve Position should be greater than the values -0.817 and -2.486, respectively, and Oil Feed minus Oil Vent Pressure, Hot Water Valve Position less than the values 0.211 and -0.008, respectively. Additionally, we can notice that in addition to the clear explainability offered by our proposed method, its F1-Score is 16.284% higher than the F1-Score of DT; meaning that our proposed method has generalized well better than DT to unseen data and achieved a reliable prediction accuracy.

Feature Abbreviation	**Description**
VC	Condenser Valve Position
VE	Evaporator Valve Position
PO net	Oil Feed minus Oil Vent Pressure
VH	Hot Water Valve Position
TO sump	Temperature of Oil in Sump
TWI	Temperature of City Water In
FWC	Flow Rate of Condenser Water
TRC sub	Liquid-line Refrigerant Subcooling from Condenser
Heat Balance%	Calculated 1st Law Energy Balance for Chiller
TO feed	Temperature of Oil Feed
TCA	Condenser Approach Temperature

Table 3.5: Attributes' abbreviations and their description for the chiller datasets

3.3.2 Occupancy Estimation

Occupant behaviour has been a central interest for research efforts [178]. The majority of research papers in this field have focused on finding the representation of the diversity of occupants' behaviour through statistics [179–181]. Additionally, recent efforts have attempted to discover the impact of specific occupants' actions on energy consumption [182]. Human behaviour changes periodically, and so does its impact on energy consumption. The application of occupancy estimation aims to:

(1) Make a discrimination between what is known as the occupant's irregular energy consumption and the true consumption that is usually incurred by the building.

(2) Design power demand response programs that rely on human activities for planning.

The energy efficiency of residential buildings can be improved through the automatic adjustment of the building's energy usage through occupancy estimation [183–186]. Several research papers have proposed a variety of applications that make use of occupancy data to handle novel building management optimizations. As an example, a management system can reduce energy consumption in response to the estimated occupancy by disconnecting inactive appliances [187], adjusting the set points for air conditioners [188] and dimming lights [189]. A key requirement for the applications that make use of occupancy data is reliability, low cost and non-intrusiveness. Using GPS information through smartphones to monitor occupancy was not an effective measure since it needed continuous and active participation by the occupants. Consequently, this solution has caused unreliability and privacy concerns. Therefore, several research efforts have attempted to deploy environmental sensors to collect data for occupancy estimation applications [184, 190]. However, several environmental sensors can be unreliable. As an example, motion detectors must be continuously calibrated and strategically positioned as they might register false occupancy or be triggered by pets [191]. Sensors that register the concentration of CO_2 in the room might be affected by external factors [192]. Additionally, power demand reduction strategies could extrapolate well if it was designed based on the least intrusive means of observing the occupancy count. Thus, in this chapter, we have decided to tackle this issue by proposing a semi-supervised method to estimate the number of occupants within a building, given the timestamp, the power demand, and the outdoor temperature. We alter these timestamp values using a standard trigonometric transformation into a two-dimensional variable that is distinguishable and distance consistent as follows:

$$(X_{i(D+1)}, X_{i(D+2)}) = (\cos(X_{i(t)}), \sin(X_{i(t)})) \tag{91}$$

where $X_{i(t)}$ represents the time stamp for observation i that is normalized between 0 and 2π. The term $(X_{i(D+1)}, X_{i(D+2)})$ is the resulting two-dimensional unit vector on the circle. Furthermore, we have considered explainability to demonstrate the factors that define the boundaries between the detected patterns that resemble the different occupancy estimates. Thus, the important factors that indicate occupancy are exhibited. Additionally, in this application, we propose a framework that takes advantage of the integration of explainability with our proposed model to maximize the pattern recognition accuracy and enhance the prediction interpretability as demonstrated in Figure 3.5. The low-level features, such as the two-dimensional unit vector calculated using Equation 91 are used in training the data-driven model, while the high-level timestamp, in addition to the labels obtained from the data-driven, is used to train the decision tree. The dataset used in this section is collected from an office in the Grenoble Institute of Technology [178, 193]. The original dataset has nine features, and the observations are recorded every 10 minutes for ten days. The original set of features are listed as follows: Time, the concentration of carbon dioxide inside the office, the office power demand in Watts, readings from a motion detector, building corridor carbon dioxide concentration, temperature inside the office in Celsius, outdoors temperature in Celsius, door open/closed status, root mean square pressure of sound and window open/closed status. However, as mentioned earlier, we have only used non-intrusive attributes, and they are listed as follows: Time, Building power demand and outdoor temperature. Our proposed model achieves an accuracy score of 91.77% in terms of estimating the correct number of occupants in the buildings shown in Table 3.6. As mentioned earlier, our model can generalize better than supervised models like DT. Thus, the explainability offered by our proposed model is more credible; that can be exhibited using the F1-Score within table 3.6. In comparison to the classification performance of the DT, our proposed model achieves an F1-Score that is 16.25% higher. F1-Score is the average harmonic mean of the precision and recall scores. Both precision and recall scores emphasize the model's ability to not classify any observation with a specific class with a different class and contribute equally to the F1-Score. The proposed model achieves this score given non-intrusive features and the assumption that only 10% of the available training data is labelled. Additionally, the explainability integrated within our proposed method allows us to define the boundaries between the detected patterns and presents the conditions to estimate the correct count of occupants in the building given a subset of the features with the minimum amount of mistakes, as shown in Figure 3.12. Each If-Then statement in the previously mentioned figure indicates the threshold value of specific features and the corresponding estimation of the number of occupants present in the building. For example, if the power attribute value is less than 1.226, greater than 1.226 or greater than 1.731, then there are 0,1 and 2 occupants in the building, respectively.

In our semi-supervised approach, where labels are available for all the datasets, we first divide the dataset evenly into training and testing subsets. The labels are assumed to be unknown in the testing and the unlabelled training subsets.

3.3.3 Energy consumption characterization

The behaviour of the energy consumption is the most important factor that characterizes the energy consumption of a building. Suitable consumption incentives can be planned by utility companies if consumer behaviour is correctly modelled and characterized. Thanks to the fast-evolving energy metering devices, there are several publicly available datasets that record the total power demand of not only the building but the power demand of several electrical appliances inside it in an independent manner [194]. Thus, the characterization of residents' energy consumption using data-driven models has increased in popularity. In this application, we're characterizing the consumption behaviour of residents based on their aggregate power demand in addition to the power demand of individual appliances.

The RIFIT dataset

The dataset used in this experiment was recorded within a project that was done in collaboration among the Universities of Strathclyde, Loughborough and East Anglia [195]. The project is entitled Personalised Retrofit Decision Support Tools for UK Homes using Smart Home Technology (REFIT), which was done in support of the Engineering and Physical Sciences Research Council (EPSRC). The data was collected from 20 households across the years 2013 and 2014. The records collected from each household consist of time-series data from 10 power sensors that consist of a current clamp and 9 Individual Appliance Monitors (IAMs). The power demand is collected in unit Watts with a time interval of 6-8 seconds.

Furthermore, we extract several statistical features that improve the clustering performance [73] and provide an interpretable representation of the load curve for each household. In order to summarize the information included in the load curve of each energy consumption over the years, we have extracted nine features. Following [73], seven features are extracted after the definition of four key time periods, and they are denoted by $t \in \{1, 2, 3, 4\}$. The **Overnight** time period ($t = 1$) is defined between 10:30 PM and 6:30 AM, the **Breakfast** time period ($t = 2$) is defined between 6:30 AM and 9:00 AM, the **Daytime** period ($t = 3$) is defined between 9:00 AM and 3:30 PM, the **Evening** time period ($t = 4$) is defined between 3:30 PM and 10:30 PM. Based on the four previously explained prominent time periods, seven features are extracted from smart meter records to summarize the representation of energy consumers, and they are calculated as follows:

- RAP_t denotes the **Relative Average Power** within the time period (t) over the entire year; it is defined as follows:

$$\textbf{RAP}_t = \frac{\text{AP}_t}{\text{DAP}}, t = 1, 2, 3, 4 \tag{92}$$

- the Mean STD denotes the **Mean Relative Standard Deviation** of the average power used over the entire year; it is defined as follows:

$$\textbf{Mean STD} = \frac{1}{4}\sum_{t=1}^{4} \frac{\sigma_t}{\text{AP}_t} \tag{93}$$

- The **seasonal score** is defined as follows:

$$\textbf{Seasonal Score} = \sum_{t=1}^{4} \frac{|AP_t^W - AP_t^S|}{AP_t} \tag{94}$$

- The **Weekend vs Weekday Difference Score** (WD-WE diff. Score) is calculated as follows:

$$\textbf{WD-WE diff. Score} = \sum_{i=1}^{4} \frac{|\text{AP}_t^{\text{WD}} - \text{AP}_t^{\text{WE}}|}{\text{AP}_t} \tag{95}$$

- the Mean STD_a denotes the **Mean Relative Standard Deviation** of the average power used over the entire year for each appliance; it is defined as follows:

$$\textbf{Mean STD}_{\alpha} = \frac{1}{4}\sum_{t=1}^{4}\frac{\sigma_{\alpha t}}{\mathrm{AP}_{\alpha t}} \tag{96}$$

where $\alpha \in \{1, 2, 3, ..., 9\}$, AP_t, and σ_t represent the average power used by the specific consumer and its corresponding standard deviation in the time period (t) respectively over all the available smart meter records data. Additionally, $\mathrm{AP}_{\alpha t}$, and $\sigma_{\alpha t}$ denote the average power demand and its corresponding standard deviation, respectively, of appliance α within the time period t. DAP represents the average daily power used by the specific consumer throughout the available smart meter data. AP_t^W and AP_t^S represent the average power used by the specific consumer in the time period (t) throughout winter and summer, respectively. $\mathrm{AP}_t^{\mathrm{WD}}$, and $\mathrm{AP}_t^{\mathrm{WE}}$ represent the average power used by the specific consumer in the time period (t) throughout the weekdays and weekends respectively for the available data. Finally, we use the proposed model to explore the hidden patterns within the training data in an unsupervised manner since the data is unlabelled. We first use the MML model selection criterion to discover the true number of consumption patterns within the training data by minimizing the objective function formulated in Equation 74. As a result, the optimal number of intrinsic groups is concluded to be four within the dataset used in this experiment. According to several clustering indexes, as shown in Table 3.7, our proposed model achieves the best performance in comparison with several clustering algorithms. Similar to the previous experiments, we provide an interpretation of the detected patterns using the integrated DT as shown in Figure 3.13. As demonstrated in the previously mentioned figure, the If-Then statements highlight the boundaries between the detected four energy consumption patterns labelled using the range [0,3]. As demonstrated in Figure 3.13, we can conclude that each energy consumption profile was characterized by specific attribute values by our proposed method as follows:

- Consumption profile 0 has an average power demand mean relative standard deviation that is less than the value of 1.365.

- Consumption profile 1 has a Weekend vs Weekday Difference Score greater than 0.774.

- Consumption profile 2 has an annual relative average power demand for breakfast time greater than the value of 0.781.

- Consumption profile 3 has a Mean Relative Standard Deviation that is greater than 1.365.

Figure 3.11 presents the detected patterns and their daily average normalized power demand. As demonstrated, it is difficult to characterize several consumption patterns visually, except for the energy consumption profile 2, which can be distinguished by its high power demand relative to the other clusters at the breakfast time period (t = 2), which is also demonstrated in Figure 3.13.

Additional conclusions could be drawn from Figure 3.11. Households following the consumption profile 1 have a noticeably higher energy consumption, which may indicate a larger housing for consumers classified within this cluster. The categorization with consumption profile three may indicate a household whose occupants work outside the house during the day, for there is exclusively high consumption during the breakfast period and relatively lower consumption during the rest of the day. The categorization under consumption profile two may indicate that the occupants work at home during the overnight time period.

Performance Measure	ExBAGGU	ExGU	ExAdaboost	ExKNN	DT
Accuracy	91.778	88.611	88.167	86.389	86.278
Silhouette	0.059	0.056	-0.028	0.065	0.052
Calinski Harabasz	17.452	20.355	63.711	43.721	42.198
Davies Bouldin	2.440	2.623	2.736	4.401	3.280
Fowlkes Mallows	0.737	0.732	0.72	0.698	0.698
Mathiews Correlation Coefficient	33.637	15.100	30.308	26.766	27.172
F1-Score	81.944	71.528	70.417	65.972	65.694
Jaccard	0.819	0.715	0.704	0.660	0.657
ROC AUC	0.885	0.822	0.815	0.787	0.786
V Measure	0.210	0.069	0.174	0.131	0.140
Rand	0.740	0.603	0.703	0.673	0.677
Normalized Mutual Information	0.147	0.069	0.134	0.131	0.140
Mutual Info	0.144	0.043	0.045	0.023	0.034
Homogeneity	0.149	0.048	0.160	0.136	0.148
Completeness	0.149	0.126	0.190	0.126	0.133
Adjusted Rand	0.430	0.143	0.392	0.342	0.353
Adjusted Mutual Info	0.264	0.060	0.165	0.120	0.129

Table 3.6: Performance evaluation of the models used in the comparison for the occupancy estimation application

Performance Measure	ExBAGGU	ExGU	ExKMC
Silhouette	0.616	0.189	-0.039
Calinski Harabasz	19.129	12.614	0.349
Davies Bouldin	0.579	1.238	5.356

Table 3.7: Performance evaluation of the models used in the comparison for the energy consumption categorization application

In the two subsequent experiments, We compare the performance of our proposed mixture

Performance Measure	BAGGUMM	BAGGMM	AGGMM	BAGMM	AGMM	BGMM	GMM
Silhouette	0.317	0.201	0.186	0.179	0.169	0.165	0.083
Davies-Bouldin	1.055	1.174	1.424	1.490	2.134	2.005	2.490

Table 3.8: Performance evaluation of the models utilizing the London project dataset

Performance Measure	BAGGUMM	BAGGMM	AGGMM	BAGMM	AGMM	BGMM	GMM
Silhouette	0.247	0.153	0.137	0.129	0.119	0.115	0.033
Davies-Bouldin	16.960	18.079	18.328	18.394	19.038	19.109	19.394

Table 3.9: Performance evaluation of the models utilizing the CER dataset

model against state-of-the-art mixture models within the same application, such as the bounded asymmetric generalized Gaussian mixture model (BAGGMM), the asymmetric generalized Gaussian mixture model (AGGMM), the bounded asymmetric Gaussian mixture model (BAGMM), the bounded Gaussian mixture model (BGMM), and the Gaussian mixture model (GMM).

Performance Evaluation Using London Residential Smart Meter Dataset

We used the Low Carbon London project dataset [92], with smart meter readings from 5567 homes between 2011-2014. After cleaning, 3891 observations from 2013 were analyzed. The dataset includes Dynamic Time of Use and Standard tariffs, as well as five different geo-demographic groups.

Table 3.8 shows the performance evaluation of different mixture models utilizing the London project dataset based on Silhouette and Davies-Bouldin scores.

The table shows BAGGUMM, our proposed mixture model, outperforms other models on the London project dataset with the highest Silhouette score (0.317) and lowest Davies-Bouldin score (1.055), signifying distinct, well-separated clusters. Other models have similar but inferior results, while AGMM, BGMM, and GMM perform poorly. BAGGUMM's superior scores suggest that it is an effective tool for analyzing the dataset. We demonstrate the resulting clusters using Figure 3.7.

Additionally, the integrated decision tree provides clear explainability, as demonstrated in Figure 3.10. It is used to analyze the characterization of energy users based on the ”seasonal score” and ”weekDays diffScore” features. Specifically, when the ”seasonal score” value is greater than 3.168, the observation is assigned to class 1. Observations with a ”seasonal score” greater than

1.417 belong to class 3. As for the ”weekDays diffScore” feature, a value greater than 0.936 leads to classification into class 2, while values less than or equal to 0.963 results in the assignment to class 0.

Performance Evaluation Using Irish Residential Energy Consumption Dataset

The first dataset for our study was collected by the *Commission for Energy Regulation* (CER) and was made publicly available by the *Irish Social Science Data Archive* (ISSDA) [89]. The dataset contains smart meter records of more than 6000 Irish energy consumers, which were collected from July 14, 2009, to December 31, 2010. The energy consumption is measured in kilowatt-hours (kWh) with an interval of 30 minutes. There are two types of energy consumers in this dataset: residential and small to medium enterprises. However, our study is focused only on residential energy consumers. Therefore, after cleaning the data, we have 3639 Irish residential energy consumers for analysis. Each residential consumer in this dataset is assigned six different tariffs, denoted by (E, A, D, C, B, and W).

BAGGUMM, our proposed mixture model, excels with a Silhouette score of 0.247 and Davies-Bouldin score of 16.960, indicating well-separated, distinct clusters. Other models, BAGGMM, AGGMM, and BAGMM, perform similarly but worse, while AGMM, BGMM, and GMM have the poorest performance. Overall, BAGGUMM outperforms all models in clustering the CER dataset. The resulting clusters are demonstrated using Figure 3.6 As shown in Figure 3.9, the tree is easily interpretable, with the ”meanSTD” and ”weekDays diffScore” features guiding the classification. Observations are assigned to class 2, class 0, or class 1 based on specific threshold values for these features.

3.4 Conclusion

In this chapter, we have proposed a statistical model that is a mixture of mixtures of bounded asymmetric generalized Gaussian and Uniform distributions. In order to learn the proposed model’s parameters using the training data, we have proposed an unsupervised and semi-supervised learning approach using the expectation-maximization (EM) algorithm. Additionally, we have derived the objective function of the minimum message length (MML) criterion in order to discover the true number of components within the training dataset for the unsupervised learning approach. In other words, our approach assumes that the data arise from a mixture of mixtures of bounded asymmetric generalized Gaussian and Uniform distributions. After the model training is done, the testing data are labelled using the trained model parameters and Bayes' rule. Three real-life applications and their datasets have been used to validate the proposed model and compare its performance to several state-of-the-art machine learning models. The final results demonstrated that the proposed model outperforms all the state-of-the-art machine learning models used in the comparisons in both the semi-supervised and the unsupervised learning tasks.

Prior works in the chiller fault detection and diagnosis application do not extrapolate as well as our proposed model, especially in the case of the scarcity of data. The bounded asymmetric generalized Gaussian distribution is flexible to fit several data distributions that may follow a density distribution that can be asymmetric and have a bounded support region. Additionally, the uniform distribution in the mixture of mixtures increased the model's robustness to outliers. Thus, it needs no prior method to remove the outliers before training the model. Additionally, our model extension for explainability provides an accurate diagnosis of the detected patterns in terms of the features used to train the model.

In the occupancy estimation application, our proposed model is trained with a small number of labelled data in comparison to the prior works. Nevertheless, our proposed model correctly predicts the number of occupants in the building given features that are non-intrusive and outperform all the state-of-the-art machine learning algorithms used in the comparison. Additionally, within this application, we propose a novel workflow that attempts to learn better models using low-level features and still provide explainability using high-level features.

In the experiment within the energy consumption categorization application, the proposed model clusters the households based on their aggregate power demand and the power demand of several individual appliances and outperforms several state-of-the-art clustering algorithms. Hence, our approach to analyzing real-life energy consumption data is effective in determining households that are suitable for demand reduction initiatives such as DR and EE. Thus, providing the opportunity for utility companies to adopt environmentally friendly and cost-effective technologies. Additionally, we extend the proposed mixture model to offer explainability and define the boundaries between the discovered energy consumption profiles in terms of statistically extracted features. The boundaries also help utility companies lay more accurate plans to motivate energy consumers to migrate to a cluster of energy consumers that consume less energy or to a cluster with a lower impact on energy sources.

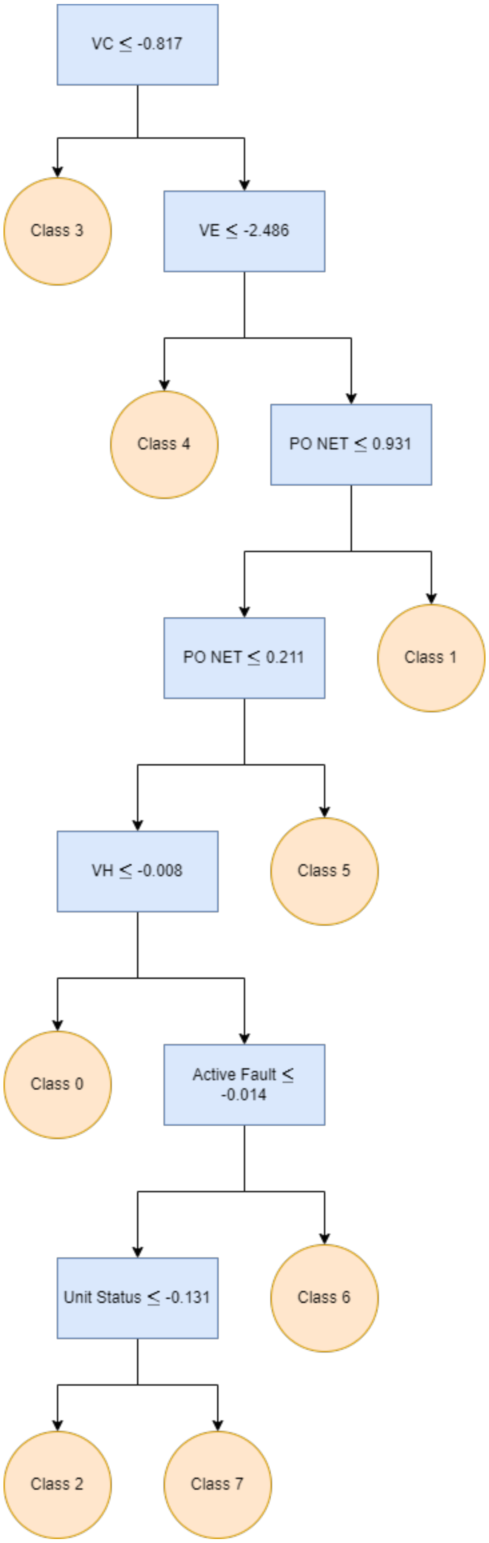


Figure 3.1: Fault severity of level 1

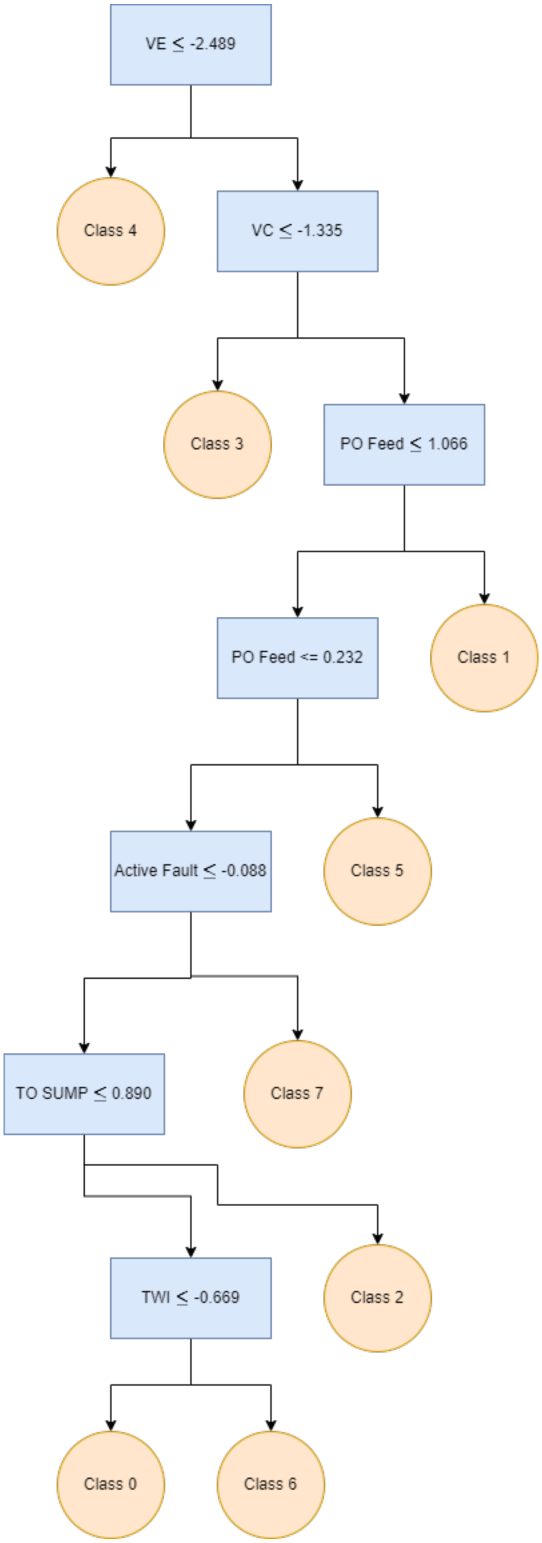


Figure 3.2: Fault severity of level 2

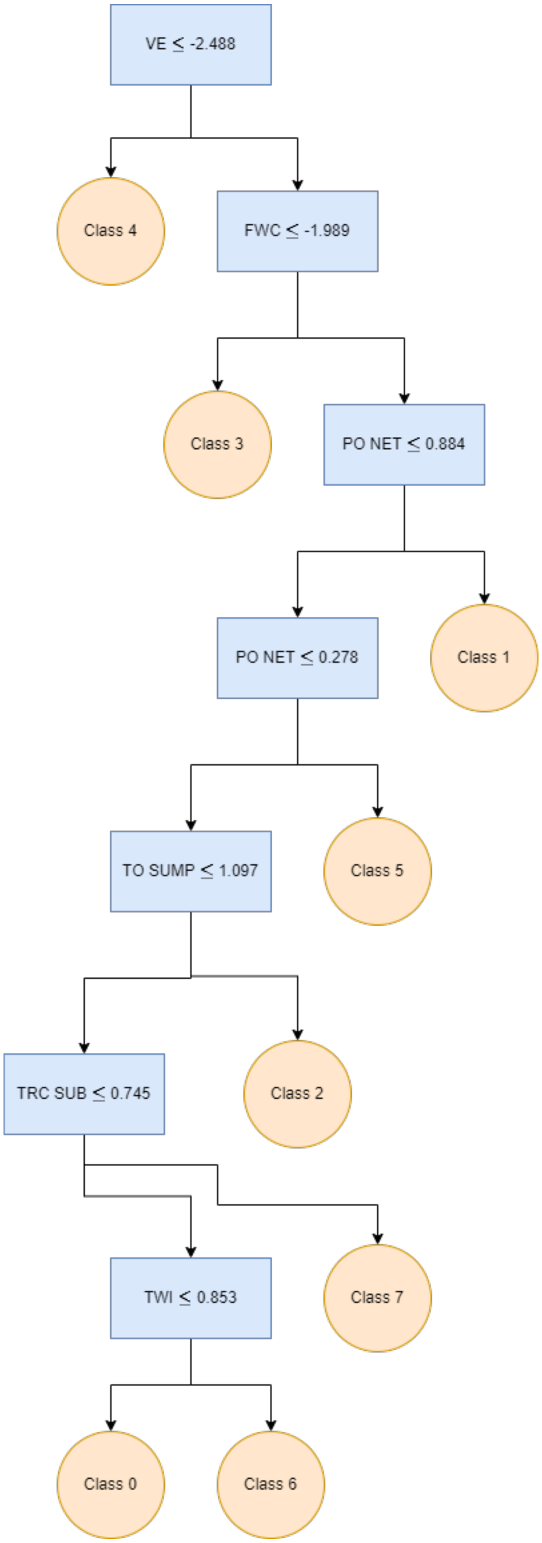


Figure 3.3: Fault severity of level 3

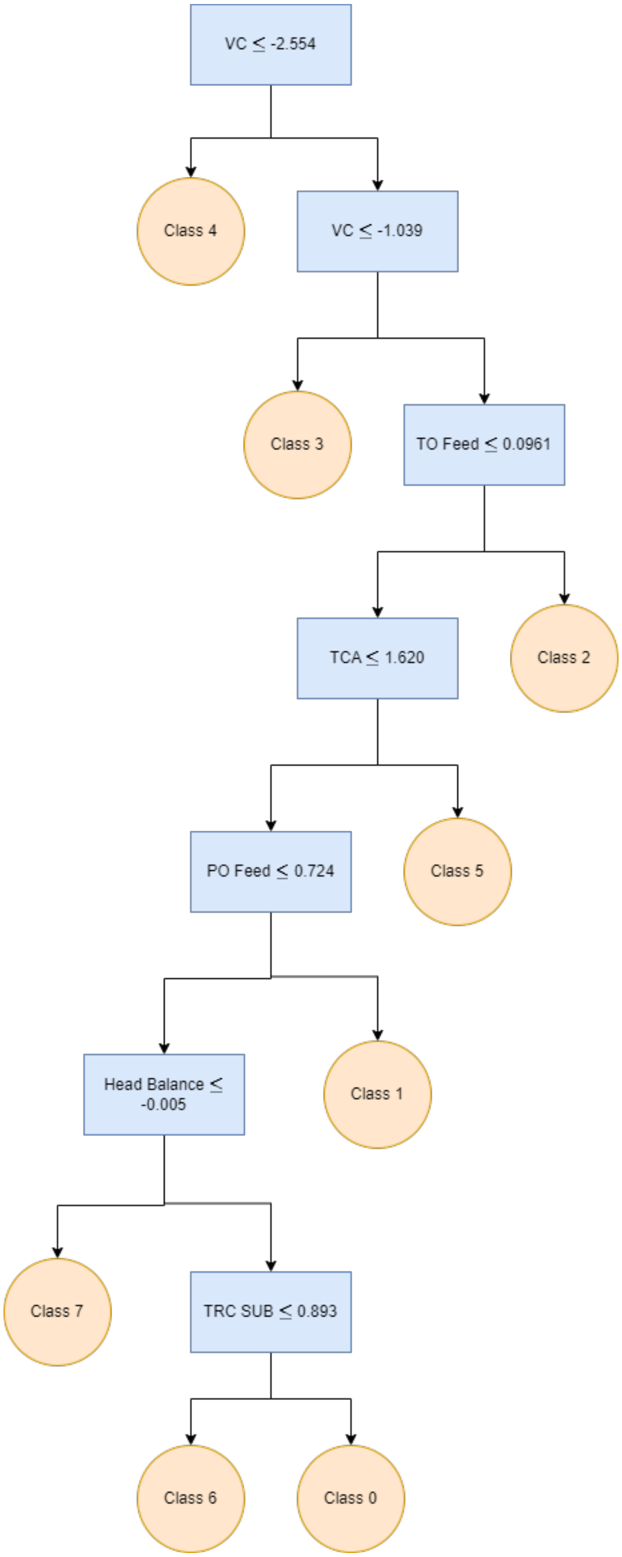


Figure 3.4: Fault severity of level 4

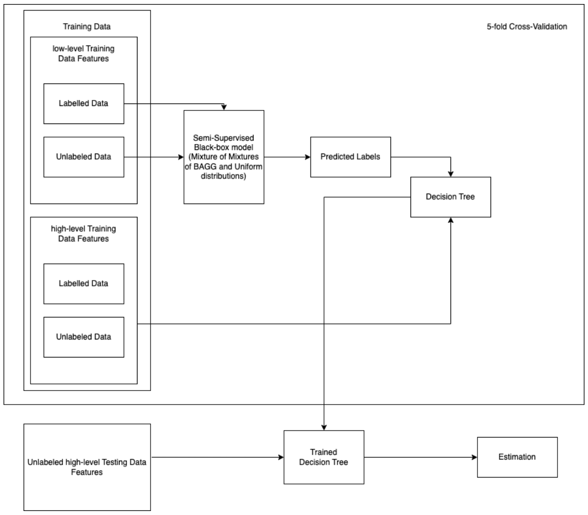

Training Data
5-fold Cross-Validation
low-level Training Data Features
Labelled Data
Unlabeled Data
Semi-Supervised Black-box model (Mixture of Mixtures of BAGG and Uniform distributions)
Predicted Labels
Decision Tree
high-level Training Data Features
Labelled Data
Unlabeled Data
Unlabeled high-level Testing Data Features
Trained Decision Tree
Estimation

Figure 3.5: The proposed workflow

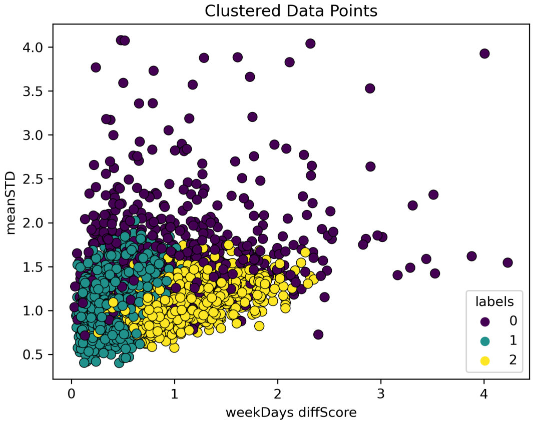


Figure 3.6: CER Clusters

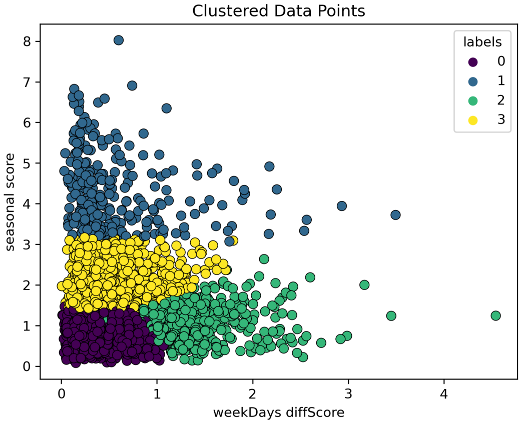


Figure 3.7: London smart meter data clusters

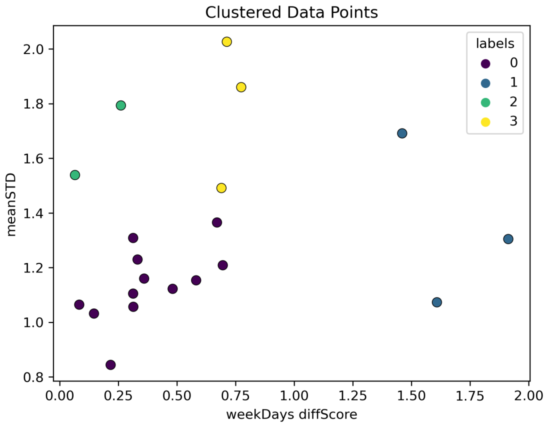


Figure 3.8: RIFIT clusters

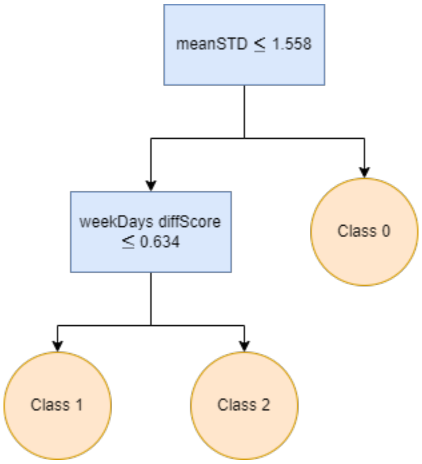


Figure 3.9: If-Then explainability for the CER dataset

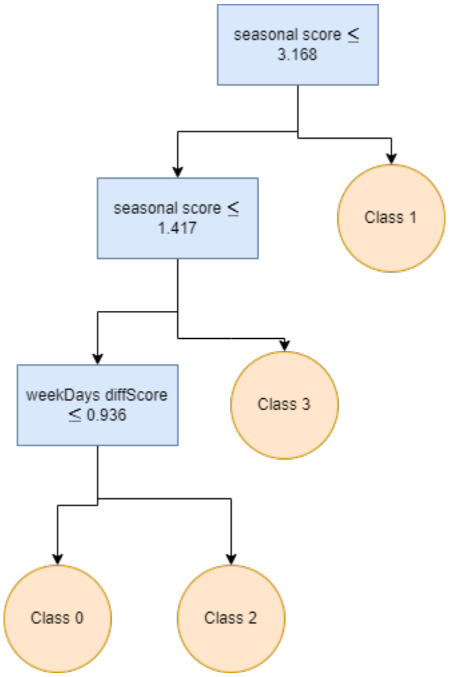


Figure 3.10: If-Then explainability for the London residential smart meter dataset

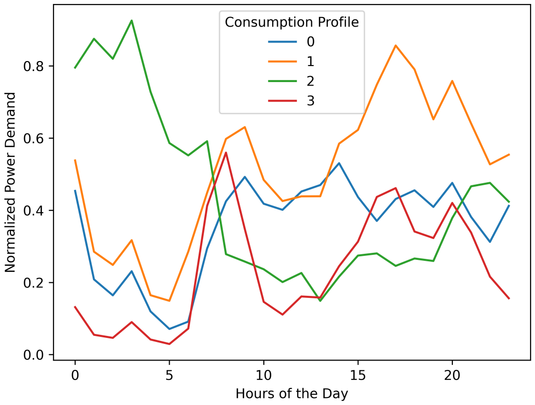


Figure 3.11: Average daily profiles of the learned patterns

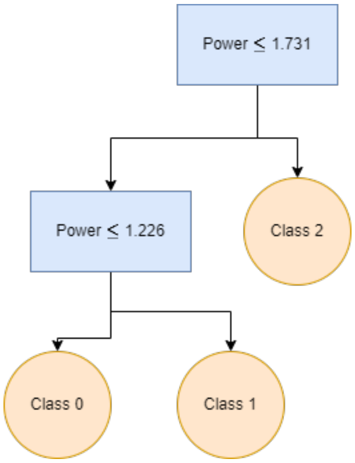


Figure 3.12: The proposed model's occupancy estimation explainability in terms of power demand

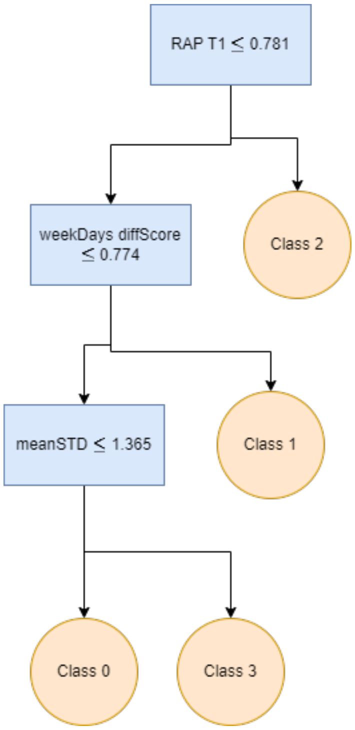


Figure 3.13: The proposed model's energy consumption categorization explainability in terms of the extracted statistical features

Chapter 4

Advanced Models for Human Activity Recognition using Mixture-Based Hidden Markov Models with Feature Saliencies

The recognition of human actions is an important research area. Recognizing human actions from a sequence of observations can be a complex task requiring an equally complex system. As a solution, in this chapter, we propose an asymmetric generalized Gaussian mixture-based hidden Markov model (AGGM-HMM) and a bounded asymmetric generalized Gaussian mixture-based hidden Markov model (BAGGM-HMM). In the previous chapter, we have been able to explore the potential of incorporating the bounded asymmetric generalized Gaussian distribution in mixture models. The resulting proposed models have been proven successful in several real-life applications and have outperformed several state-of-the-art models in the context of real-life applications. However, the temporal domain was a component we did not consider during our previous pattern recognition processes. Therefore, in this chapter, following our introduction of hidden Markov models (HMMs) in Section 1.5, the potential of mixture-based HMMs incorporating AGGMM and BAGGMM is explored by integrating them within HMMs. For instance, BAGGM-HMM adopts

BAGGMM to model emission probabilities, while AGGM-HMM adopts AGGMM for the same purpose.

In addition to the aforementioned proposed models, a joint feature selection and parameter estimation approach is developed for all the proposed mixture-based hidden Markov models (MM-HMM) within this chapter (AGGM-FSHMM and BAGGM-FSHMM). The asymmetric generalized Gaussian distribution (AGGD) is a flexible distribution robust to outliers and is capable of modelling densities that are asymmetric and non-Gaussian. The bounded variant of the AGGD (BAGGD) further enables the distribution to better model real-life applications' data. As an additional parameter set defining the MM-HMM, feature saliencies enable the proposed model to obtain additional discrimination power between states. We use the Expectation-Maximization (EM) algorithm for all the proposed models to obtain the maximum likelihood point estimates of the complete set of parameters. This chapter has utilized two approaches to the human activity recognition (HAR) application to validate the proposed frameworks: sensor-based HAR and video-based HAR. We use raw data recorded from sensors attached to the subjects for sensor-based HAR. For video-based HAR, we leverage histograms of optical flow (HOF) and motion boundary histogram (MBH) descriptors for the unsupervised pattern recognition process. Seven real-life data sets are utilized within the experimental part to demonstrate that our proposed frameworks are superior to several state-of-the-art HAR methods.

4.1 Introduction

The use of HMMs has been highly successful in various fields, such as speech recognition and image categorization. This success has been achieved through the introduction of new methods that work together with HMMs to enhance their performance, as well as through the analysis of a wide range of data types and features. Hidden Markov models (HMMs) are an extension of Markov models and share a similar property in that the state sequence of an HMM forms a Markov chain. However, HMMs are more general as they allow for observations to move between states over time, which is not the case in simple Markov models. While mixture models assume that each observation is independent, HMMs assume dependence between observations. This is why [196] referred to HMMs as dependent mixture models. The sequence of components in an HMM from which observations are drawn also follows a Markov process, meaning that consecutive components are related through a Markov chain. Both Markov models and HMMs have finite states, with the states in HMMs representing elements from a finite set. An HMM is a well-grounded dual stochastic model that extracts underlying statistical data via a concise feature set [78]. The primary structure of an HMM is composed of a Markov chain of latent variables, where each variable corresponds to a conditioned observation [78]. Markov chains are useful for modelling sequential patterns in time series data without the need for the independent identically distributed assumption while also maintaining generality [81].

In their daily lives, people frequently engage in a variety of routine and official activities such as driving, cleaning, and playing games. These activities require a number of fundamental actions, such as standing, sitting, bending, running, and typing. To develop a system for human-computer interaction, it is necessary to identify the specific actions that a person is performing. While several methods for recognizing human activities (HAR) have been studied for many years, the most significant developments in the field have occurred in recent decades [197]. This is particularly evident in the visible spectrum, where a large amount of data has been made accessible [198, 199]. Examples of such data include the UCF101 [200], KTH [201], and Weizmann [202] datasets.

Nonetheless, there are still several difficulties that persist in the field despite the advancements made. One such challenge is the intrinsic within-class variability, where an individual may perform the same action in a distinct manner from another individual [203]. In addition, some challenges are unique to the visible spectrum, such as the sensitivity of the visible spectrum to issues like shadows, background clutter, occlusion, and fluctuations in illumination [204].

Despite the effectiveness of employing the Gaussian-based HMM in all HMM instances, it is not the best practice to do so with all types of data. Utilizing the Gaussian distribution leads to sub-optimal modelling due to the fact that the distribution has unbounded, infinite support, a rigid bell shape, and symmetry properties. Using probability density functions that are better suited to the data of an application as an emission (probability density functions) PDFs has been demonstrated to be effective in improving the performance of a hidden Markov model (HMM) [205–208].

This expansion of data has greatly contributed to the advancement of machine learning techniques. However, it has also presented a number of difficulties and challenges, including the challenge of managing high-dimensional data, which can result in the curse of dimensionality. The literature on HMMs typically concentrates on discrete- and Gaussian-based HMMs [79]. However, there has been a recent effort to improve the modelling of state emission probabilities by taking into account the characteristics of the data [209, 210].

One common method for learning HMMs is the expectation-maximization (EM) algorithm, which involves iteratively computing the expected values of the latent variables and maximizing the likelihood of the observed data. However, EM can be computationally expensive, and it may not converge to a global optimum. To address these issues, a variation of the EM algorithm called the Baum-Welch algorithm is often employed. This algorithm uses the forward-backward algorithm to compute the expected sufficient statistics of the latent variables, which are then used to update the model parameters. The Baum-Welch algorithm has several advantages over the standard EM algorithm, including faster convergence, robustness to initialization, and the ability to handle missing data.

The several contributions within this chapter include mixture-based hidden Markov models. The theoretically assumed model to present the best performance assumes that the underlying distribution of the observations at a specific time within the sequence follows a bounded asymmetric generalized Gaussian mixture model, and the underlying state sequence follows a Markov process. Adopting the bounded asymmetric generalized Gaussian mixture model within the proposed MM-HMM has several advantages over adopting a single probability density function [211], and they are:

- Improved modelling of feature distributions: The BAGGM-HMM model allows for more complex modelling of the distribution of feature vectors by using a mixture of asymmetric generalized Gaussian mixture to model each hidden state. The proposed model allows for more flexibility in modelling the underlying distribution of features, which can be important for complex actions involving multiple sub-actions or movement variations.

- Better handling of feature variability: The BAGGMM-HMM model can handle variability in the observed features more effectively by allowing each hidden state to have a different mixture model. The proposed model allows the model to capture different modes of variation in the feature space, which can be important for recognizing subtle differences in actions.

- Reduced overfitting: The BAGGM-HMM model is less prone to overfitting than the single distribution-based HMM model since it has more parameters and, thus, more flexibility to model the data. This property can be important when working with small datasets, which are common in the context of HAR.

In addition, we acknowledge the possibility of the existence of noisy and uninformative features. Consequently, we approximate a group of parameters defining the most favourable feature subset, enhancing the model's parameter set optimization process. Feature selection within our proposed models is accomplished by expanding on the feature saliency model introduced by [212].

Overall, the chapter introduces the mathematical model for maximum likelihood estimation learning of AGGMM-HMM, BAGGMM-HMM, AGGM-FSHMM, and BAGGM-FSHMM, and employs these models to evaluate various HAR datasets with different modalities, such as sensor-based and video-based datasets. The purpose of this evaluation is to investigate the generalization capabilities of the proposed models. As far as we are aware, this is the first instance where AGGM-based HMMs are being assessed in both video-based and sensor-based HAR applications.

The structure of this chapter is as follows: Section 4.2 provides a summary of prior studies that have utilized HMMs, and we discuss the applications of HMMs in conjunction with general feature selection methods as a predictive model. In Section 4.3, we explain some background information that is necessary for the presentation of the proposed models. In Section 4.4, we introduce the mixture models used within the proposed HAR frameworks. Section 4.5 introduces how we integrate mixture models within the proposed HMMs. Section 4.6 introduces how we integrate feature selection within the proposed HMMs. Section 4.8 presents the experimentation and analysis of results pertaining to real-life problems. In section 4.9, the chapter concludes with a summary of the work and concluding remarks.

4.2 Related work

4.2.1 Hidden Markov models

In the field of HAR, two primary types of machine learning techniques are commonly employed for data modelling: discriminative and generative, as noted in [213]. The discriminative approach involves establishing a relationship between input and output data, while the generative approach assumes that the data is derived from a fundamental distribution and then applies an optimization process to identify that distribution. There exist various optimization methods for implementing the generative approach, each with its own particular strengths.

Any approach to recognizing human actions requires the extraction of informative measurements from the video and their subsequent categorization. In the last two decades, numerous investigations have been conducted in the field of human action recognition. Three primary categorization approaches have been adopted: 1) recognizing actions directly in the time domain (e.g., [214] and [215]); 2) recognizing actions using graphical models, such as HMM [216–218], dynamic Bayesian networks, and conditional random fields [219–222] recognizing actions using histograms of measurements [201, 220, 223].

The time domain approach uses dynamic time warping (DTW) as its primary representative [214]. Graphical models, specifically HMM, are effective in recognizing actions [216, 217, 224], with other models such as dynamic Bayesian networks and conditional random fields also being successful [221, 222]. While histogram-based approaches have shown significant empirical accuracy [201, 223, 224], they do not fully account for the temporal dimension of human actions.

Sequential classifiers like HMM can naturally classify sequences of any length and have demonstrated an ability to adjust to variations in the duration of instances of the same action. HMMs can be easily incorporated into complex hierarchical models, as demonstrated in [225]. Therefore, the following discussion will focus on HMM and provide extensive comparisons with other approaches.

4.2.2 Feature selection

The field of feature selection is extensive and has seen the implementation of several techniques to reduce the number of features in both supervised and unsupervised scenarios, as noted in [226].

As per [212], feature selection methods are generally classified into three primary categories in current research, which are filters, wrappers, and embedded techniques. Additionally, meta-heuristic methods are also included in these categories.

The process of selecting a relevant subset of features for model building can be done using filter methods like information gain [227, 228], Pearson's correlation coefficient [229], and variance threshold [72]. These methods evaluate all the features and return the most relevant subset. However, wrapper methods focus on optimizing the classifier's performance and are more involved in the model-building process.

The identification of relevant features using wrappers is a common practice in machine learning. The selection methods usually used in wrappers are forward selection, backward elimination, or recursive feature elimination, as cited in [230], [231], [232], [233], and [234]. Wrappers build the model using a subset of features depending on the learning algorithm and evaluate its performance using specific criteria. In other words, wrappers measure the performance of the model based on a particular set of features selected by the algorithm.

Feature selection is approached as an optimization problem through methods that use meta-heuristic algorithms. Evolution-based algorithms [235] can find the best solution by being flexible, simple, and avoiding local optima [236]. To begin the feature selection process, these methods generate random solutions that do not require complex derivative calculations. Then, they explore the search space to identify promising areas through a thorough investigation.

It is evident that techniques for transformation, including Principal Component Analysis (PCA) and Independent Component Analysis (ICA), effectively decrease the number of features in a model. Due to this reason, these techniques have been incorporated into Hidden Markov Models (HMMs) as reported in [237, 238].

The significance of feature selection in HMMs is primarily due to the essential requirement of identifying the appropriate feature to utilize in the model. Although there have been several studies conducted in general and in the context of mixture models in particular, the methods for feature selection in HMMs are restricted. According to to [239] and [62], there is a scarcity of dedicated feature selection techniques for HMMs. Typically, in most applications, features are selected based on prior domain knowledge, and there is a complete absence of a consistent feature selection process, as mentioned in [240] and [241].

The central theme of this chapter is embedded or integrated feature selection techniques, which involve the simultaneous consideration of all the features in a dataset. These features are used as inputs for a machine learning algorithm that is designed to optimize the performance of the model. The resulting output of this process includes both a reduced set of features and the parameters of the model. Consequently, the embedded feature selection method combines the benefits of both wrapper and filter methods, which respectively involve selecting subsets of features specific to a learning algorithm [242].

4.3 Background

The Gaussian distribution as an isotropic probability density function can be used effectively within a clustering algorithm to compactly model and represent the intrinsic grouping of the data. The representation of the Gaussian mixture model comprises a set of parameters that would not relatively yield a computationally expensive model as the dimensions of the data grow. The set of parameters of each mixture component describes each discovered pattern's properties and includes a mean parameter vector and a covariance matrix. Given this motivation, the EM algorithm has been used to efficiently estimate the parameters of a Gaussian mixture model that best fits the data in several applications [19,20]. Although this compact representation may yield clustering algorithms that require a relatively low computational cost, in some applications, it could introduce several limitations to modelling the data, such as:

(1) It has a rigid bell shape and a tail that is too short for most real-world problems [21].

(2) The Gaussian distribution and several other choices of distributions for a given mixture model are unbounded with a support range that extends from $-\infty$ to ∞ [243].

(3) The distribution is symmetric around its mean.

The data in real-life applications are usually bounded and most likely have a density that is non-Gaussian [24]. The term 'non-Gaussian' in the context of describing a distribution signifies a density function that is asymmetric, non-bell shaped, or both. The fixed kurtosis of the Gaussian distribution makes the mixture model vulnerable to the outliers of individual data clusters. Since the Gaussian distribution does not have a parameter that controls the distribution's tails to be correctly estimated while fitting the distribution to each data cluster, the distribution is unsuitable for assigning a relatively low probability of occurrence to the individual class outliers [28]. The pattern recognition applications that integrate the Gaussian mixture model require pre-processing to eliminate outliers, resulting in an increase in computational complexity [30]. On the other hand, there have been several attempts to use more flexible distributions than the Gaussian to fit data from diverse applications in addition to its ability to generalize to it. Several research papers have proposed using the generalized Gaussian distribution (GGD) in diverse real-life applications [151, 152]. The distribution is formalized previously in Equation 9. where $A(\lambda) = \left[\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}\right]^{\lambda/2}$, and X denotes the random variable. The parameters μ and σ denote the mean and the standard deviation, respectively. The symbol λ denotes the parameter that distinguishes this distribution from the Gaussian distribution. The λ parameter obeys the condition $\lambda \geq 1$ and controls the kurtosis of the distribution, which determines if the probability density function is peaked or flat. As the parameter λ increases in value, the distribution becomes flattered until it degrades to its special case of the Uniform distribution as $\lambda \to \infty$. As the λ decreases in value, the probability density function becomes more peaked. As $\lambda \to 0$, the probability density function becomes a delta function with an infinite value at μ. Practically, the λ parameter controls the tails of the distribution, and if estimated correctly for the training data, the distribution should model the data accurately with robustness to outliers. The distribution is reduced to the Gaussian distribution when $\lambda = 2$ and to the Laplacian distribution when $\lambda = 1$. Thus, the distribution is flexible and able to fit diverse real-life application data better in comparison to the use of each of its special cases individually in a mixture model. This distribution is especially effective if the data features are assumed to be distributed independently.

Furthermore, to fit asymmetrically distributed data, researchers have proposed the usage of the asymmetric generalized Gaussian distribution (AGGD) in several applications [123, 155]. The AGGD is formalized similarly to Equation 10. In comparison to the GGD, this distribution has two parameters to describe the variance of the data instead of one parameter. The left and right standard deviations are denoted as σ_l and σ_r, respectively; they control the asymmetry of the probability density function. The distribution is skewed left if $\sigma_l < \sigma_r$, skewed right if $\sigma_l > \sigma_r$ and symmetric if $\sigma_l = \sigma_r$. Thus, in addition to the flexibility provided by the GGD, the AGGD is capable of modelling data with an asymmetrical distribution [156].

4.4 Bounded asymmetric generalized Gaussian mixture model

In this chapter, we assume that the D-dimensional random vector that represents the data follows an M-component mixture model. The corresponding probability density function is formalized similarly to Equation 1. where p_m is the mixing weight for cluster m and it is governed by the following constraints: $p_m \geq 0$, $\sum_{m=1}^{M} p_m = 1$. The symbol Θ denotes the mixture's complete set of parameters. In this section, we demonstrate how the bounded asymmetric generalized Gaussian distribution (BAGGD) is adopted as a component density for the proposed mixture model. The bounded component distribution is defined in terms of the unbounded component's density distribution $\Psi(X_{id}|\theta_{md})$ following [244] and similar to Equation 8. where $\Psi(X_{nd}|\theta_{md})$ denotes the unbounded asymmetric generalized Gaussian probability density function. The terms $H(X_{nd}|m)$ and $\int_{\partial_m} \Psi(X|\theta_{md}) d\mathcal{X}$ contribute to making the AGGD used in the proposed mixture bounded. The bounded support region is denoted by τ_{md} for each component m and dimension d. The indicator function $H(X_{nd}|m)$ sets the density value of the AGGD outside the bounded support region to zero, and it is defined similarly to Equation 7. The term $\int_{\tau_{md}} \Psi(X|\theta_{md}) d\mathcal{X}$ denotes the normalization constant that restores the statistical properties of the probability density function, the share of $\Psi(X_{nd}|\theta_{md})$ that falls within the support region τ_{md}.

Additionally, we propose another mixture-based HMM in this chapter for which the emission probability is modelled by an asymmetric generalized Gaussian mixture model. The corresponding probability density function is formalized as follows:

$$p(\vec{X}|\Xi) = \sum_{m=1}^{M} \Psi(X_{nd}|\theta_{md}) p_m \tag{97}$$

where p_m is the mixing weight for cluster m. The symbol Ξ denotes the mixture's complete set of parameters. In this section, we adopt the asymmetric generalized Gaussian distribution (AGGD) as a component density for the proposed mixture model.

4.5 The mixture-based hidden Markov models

4.5.1 The hidden Markov model

For many real-life applications, predictions for a time series are made accurately if they take into account given previous sequences of observed variables. Given our consideration of the Hidden Markov Models (HMMs) as a proposed solution to the applications mentioned within this chapter, it is worth mentioning that HMMs have the Markov assumption that future predictions are dependent exclusively on the previous hidden state. The observed variables are denoted by $\mathcal{X} = [\vec{X}_1, \dots, \vec{X}_N]$ and the hidden states are denoted by $\mathcal{S} = [S_1, \dots, S_N]$; $S_{ik} \in [1, K]$ and $i = [1, \dots, N]$ where K denotes the number of hidden states. The HMM is governed by several parameters, such as the transition matrix, the emission matrix, and the stationary distribution. Three main tasks are conducted during the implementation of an HMM, and they are listed as follows:

- Optimizing the HMM parameters using a training dataset.
- Scoring: this process is defined by the calculation of the joint probability of a specific scenario of a sequence given the model.
- Decoding: this process is defined by finding the optimal series of hidden states.

According to to [79], given a sequence of hidden variables that are generated by the hidden states $\mathcal{S}$, we define the transition probability matrix $A_{jk} = P(S_{ik} = 1 | S_{i-1,k} = 1)$. The elements of matrix A are governed with the following constraints: $0 \leq A_{jk} \leq 1$, and $\sum_{k=1}^{K} A_{jk} \geq 0$. As for the transition matrix, its elements are comprised of the following elements $[P(\vec{X}_i|\Theta), \dots, P(\vec{X}_N|\Theta)]$. The joint probability distribution over the hidden and observed states is defined as follows:

$$P(\mathcal{X}, \mathcal{S}|\Lambda) = P(S_1|\pi) \Bigg[\prod_{i=2}^{N} P(S_i|S_{i-1}, A) \Bigg] \prod_{i=1}^{N} P(\vec{X}_i|\Theta) \tag{98}$$

where $\Lambda = \{\pi, \Theta, A\}$ denotes the set of parameters that characterize the proposed HMM. Considering that the observed sequences are continuous variables, previous studies have considered the

Gaussian mixture model for the emission distributions within their proposed HMMs [79, 245–247].

In this section, using the EM algorithm, we demonstrate how the maximum likelihood approach is achievable for the proposed HMM. The EM algorithm was used in order to avoid directly maximizing the likelihood function since the process is intractable.

Initial parameters are chosen at the beginning of the algorithm. Subsequently, sufficient statistics are accumulated, and the following posterior distribution $p(S|X, \Lambda^{OLD})$ is computed within the E-step. The posterior distribution is then used within the M-step to find the HMM complete set of parameters Λ. Within the EM algorithm, the M-step is achieved by maximizing the following function:

$$Q(\Lambda, \Lambda^{OLD}) = \sum_{\mathcal{S}} P(\mathcal{S}|\mathcal{X}, \Lambda^{OLD}) \log P(\mathcal{X}, \mathcal{S}|\Lambda) \tag{99}$$

Subsequently, we define the marginal posterior distribution of the nth state $\gamma(s_{nk})$ and joint posterior distribution of the present state and the state before it $\xi(s_{(i-1),j}, s_{ik})$ as follows:

$$\gamma(s_{ik}) = P(S_{ik}|\mathcal{X}, \Lambda) \tag{100}$$

$$\xi(s_{(i-1),j}, s_{ik}) = P(S_{i-1,j}, S_{ik}|\mathcal{X}, \Lambda) \tag{101}$$

The introduction of the previously defined posterior probabilities redefined our objective function as follows:

$$\begin{aligned} Q(\Lambda, \Lambda^{OLD}) = & \sum_{k=1}^{K} \gamma(s_{1k}) \log \pi_k + \sum_{i=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(s_{(i-1),j}, s_{ik}) \log A_{jk} \\ & + \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma(s_{ik}) \log P(\mathcal{X}, \mathcal{S}|\Lambda_{ik}) \end{aligned} \tag{102}$$

4.5.2 Integration of mixture models within the HMM framework

The integration of mixture models within HMM has proven effective in the context of several applications [211]. Given that the data of the applications considered within this chapter have continuous random variables within their observed sequences, mixture models such as Gaussian mixture models are considered within the HMM framework. However, considering the Gaussian distribution as a density function for each component within mixture models makes the HMM framework vulnerable to several factors. The Gaussian distribution is a bell-shaped function that is symmetric around its means and has a short tail. The symmetric form of the Gaussian function does not make it an ideal choice for asymmetrically distributed data. The fixed bell shape of the Gaussian distribution makes it vulnerable to outliers that are commonly found within class data. Therefore, within the proposed HMM framework, we propose the utilization of the asymmetric Generalized Gaussian Mixture Model. The asymmetric generalized Gaussian distribution has a shape parameter that controls its tails. The distribution also has a right and a left standard deviation that controls the skewness of the distribution. These properties make the proposed distribution robust against class outliers and able to fit data with non-Gaussian distributions even if they were not symmetric.

The integration of the BAGGMM within the proposed HMM framework is done by considering that the emission probability is represented by the BAGGMM. In the E-step, the Q function is evaluated. In the M-step, the Q function is maximized with respect to the parameters of the proposed framework to obtain the point estimates of the model's optimal parameters.

Estimation of the stationary distribution and the transition matrix

In order to obtain the optimal parameters that maximize the likelihood of using the proposed models, we derive the objective function 102 with respect to each model parameter. Therefore, maximizing the objective function with respect to the stationary distribution π and the transition matrix A yields the following point estimates:

$$\pi_k = \frac{\gamma(s_{1k})}{\sum_{j=1}^{K} \gamma(s_{1j})} \tag{103}$$

$$A_{jk} = \frac{\sum_{i=2}^{N} \xi(s_{i-1,j,s_{ik}})}{\sum_{l=1}^{K} \sum_{i=2}^{N} \xi(s_{i-1,j}, s_{il})} \tag{104}$$

Estimation of Θ

The objective function is maximized with respect to the complete set of parameters of the proposed framework. The parameter set denoted by Θ_k characterizes the kth state emission probability distribution, where $\Theta = [P_1, \dots, P_M,$

$\mu_1, \dots, \mu_M, \lambda_1, \dots, \lambda_M, \sigma_{l_1}, \dots, \sigma_{l_M}, \sigma_{r_1}, \dots, \sigma_{r_M}]$. The probability of being at state s_k at time i with respect to the mth mixture model is denoted by $\phi_n(k,m)$. Following [78, 81], $\phi_n(k,m)$ is defined as follows:

$$\rho_i(k,m) = \frac{\alpha(s_{nk})\beta(s_{nk})}{\sum_{k=1}^{K} \alpha(s_{nk})\beta(s_{nk})} . \frac{P(\vec{X}_i|\theta_{km})P_{km}}{\sum_{m=1}^{M} P(\vec{X}_i|\theta_{km})P_{km}} \tag{105}$$

where $\alpha(s_n)$ denotes the probability of the occurrence of observed variables until time N and the hidden state s_n. The symbol $\beta(s_n)$ denotes the conditional probability of the occurrence of observable variables X_{n+1} to X_N given that they are emitted using the hidden state S_n. The probabilities $\alpha(s_n)$ and $\beta(s_n)$ are formulated as follows:

$$\alpha(s_i) \equiv P(\vec{X_1}, \dots, \vec{X_n}, s_i) \tag{106}$$

$$\beta(s_i) = P(\vec{X_{n+1}}, \dots, \vec{X_N}|S_i) \tag{107}$$

The point estimate of the mixing weight P_{KM} of the mth mixture model within the kth hidden state is defined as follows:

$$P_{km} = \frac{\sum_{i=1}^{N} \rho_i(k,m)}{\sum_{i=1}^{N}\sum_{m=1}^{M} \rho_i(k,m)} \tag{108}$$

Given the fact that it is difficult to obtain a closed-form solution for several parameters of our proposed frameworks, Newton-Raphson was used in this chapter to obtain an approximate solution and demonstrated below:

$$\hat{\mu}_{kmd} = \mu_{kmd} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \mu_{kmd}^2}\right)^{-1}\left(\frac{\partial \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \mu_{kmd}}\right)\right] \tag{109}$$

$$\hat{\sigma}_{l_{kmd}} = \sigma_{l_{kmd}} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \sigma_{l_{kmd}}^2}\right)^{-1}\left(\frac{\partial \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \sigma_{l_{kmd}}}\right)\right] \tag{110}$$

$$\hat{\sigma}_{r_{kmd}} = \sigma_{r_{kmd}} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \sigma_{r_{kmd}}^2}\right)^{-1}\left(\frac{\partial \mathcal{Q}(\mathcal{X},\Lambda)}{\partial \sigma_{r_{kmd}}}\right)\right] \tag{111}$$

$$\hat{\lambda}_{kmd} = \lambda_{kmd} - \left[\left(\frac{\partial^2 \mathcal{Q}(\mathcal{X}, \Lambda)}{\partial \lambda_{kmd}^2} \right)^{-1} \left(\frac{\partial \mathcal{Q}(\mathcal{X}, \Lambda)}{\partial \lambda_{kmd}} \right) \right] \tag{112}$$

The derivatives above are calculated using the following equation considering the bounded asymmetric generalized Gaussian mixture to model the emission probabilities within the proposed BAGGM-HMM:

$$\frac{\partial \mathcal{Q}(\mathcal{X}, \Lambda)}{\partial \theta_{kmd}} = \frac{\partial}{\partial \theta_{kmd}} \sum_{i=1}^{N} \rho_i(k, m) \log(P_{km} \phi(\vec{X}_i | \theta_{kmd}) \tag{113}$$

The derivatives above are calculated using the following equation considering the asymmetric generalized Gaussian mixture to model the emission probabilities within the proposed BAGGM-HMM:

$$\frac{\partial \mathcal{Q}(\mathcal{X}, \Lambda)}{\partial \theta_{kmd}} = \frac{\partial}{\partial \theta_{kmd}} \sum_{i=1}^{N} \rho_i(k, m) \log(P_{km} \Psi(\vec{X}_i | \theta_{kmd}) \tag{114}$$

Further developments of the necessary equations needed to calculate the point estimates of the MM-HMM parameters are demonstrated in Appendix A.3.

4.6 Integration of feature selection with the mixture-based HMMs

In this section, we demonstrate how we integrate the feature saliency as parameters within the proposed BAGGM-based HMM (BAGGM-FSHMM).

4.6.1 Feature saliency within the mixture-based hidden Markov model

In this section, we shall consider an HMM with continuous emission probabilities. As denoted in previous subsections, K is the number of states. Similarly, $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_T\}$ is the sequence of observed data. Each observation at time t is denoted by $\vec{X}_t$ and $\vec{X}_t \in \mathbb{R}^D$. Naturally, X_dt denote the observation's d'th dimension at time t. The observed hidden state sequence shall be denoted by $S = \{s_1, \ldots, s_T\}$. Similar to previous sections, the transition matrix is denoted by A, and its elements are denoted as follows $a_{ij} = \{\mathbb{P}(\sim_{\approx} = \gimel | \sim_{\approx - \nVdash} = \beth)\}$. The symbol π shall denote the initial distribution or the stationary distribution. Given the above notations, we can write the complete data likelihood as follows:

$$\mathbb{P}(S, \mathcal{X}|\Lambda) = \pi_{s_1} f_{s_1}(\vec{x}_1) \prod_{t=2}^{T} a_{s_{t-1},t} f_{s_t}(\vec{x}_t) \tag{115}$$

where Λ is the complete set of parameters that defines our proposed HMM. The function $f_{s_t}(\vec{x}_t)$ denotes the emission distribution or mixture model given the state s_t.

Concerning the feature selection approach we have followed in this chapter, we have considered the feature saliency concept for mixture models representing emission distributions [62]. A feature within an observation at all times is considered relevant if it is dependent on the underlying state. The binary variable set **H=** $\{\varphi_1, \ldots, \varphi_D\}$ indicates the relevancy of each feature within all observations at all times. Subsequently, the probability of relevance of feature d is denoted by $P(\varphi_d = 1)$. In this chapter, we assume that the features are independent and the probability of $\vec{X}_t$ given S and φ is formalized as follows:

$$p(\vec{X}_t|\mathbf{H}, \Lambda, s_t = k) = \prod_{d=1}^{D} \phi(X_{dt}|\theta_{kd})^{\varphi_d} q(_{dt}|\epsilon_d, \varsigma_d^2)^{(1-\varphi_d)} \tag{116}$$

where $\phi(X_{dt}|\theta_{kd})$ denotes the bounded asymmetric generalized Gaussian distribution for the feature d, that is, for our first proposed FSHMM model. As for our secondly proposed mixture-based HMM, the corresponding probability of $\vec{X}_t$ given S and φ is formalized as follows:

$$p(\vec{X}_t|\mathbf{H}, \Lambda, s_t = k) = \prod_{d=1}^{D} \psi(X_{dt}|\theta_{kd})^{\varphi_d} q(_{dt}|\epsilon_d, \varsigma_d^2)^{(1-\varphi_d)} \tag{117}$$

As for our second proposed FSHMM model, the asymmetric generalized Gaussian distribution is considered, and it is denoted by $\psi(X_{dt}|\theta_{kd})$. The distribution $q(_{dt}|\epsilon_d, \varsigma_d^2)$ denotes the state-independent density or the background Gaussian distribution of feature d. The parameters of the latter distribution are the mean ϵ_d and the variance ς_d^2. The marginal distribution of the feature state dependence indicator set of variables H is formalized as follows:

$$P(H|\Lambda) = \prod_{d=1}^{D} P(\varphi_d = 1)^{\varphi_d}(1 - P(\varphi_d = 1))^{1-(\varphi_d)} \tag{118}$$

The joint distribution of $\vec{X}_t$ and **H** is formalized as follows:

$$P(\vec{X}_t, \mathbf{H}|s_t = k, \Lambda) = \prod_{d=1}^{D}[P(\varphi_d = 1)\phi(X_{dt}|\theta_{kd})]^{\varphi_d}[(1 - P(\varphi_d = 1))q(_{dt}|\epsilon_d, \varsigma_d^2)]^{1-phi_d} \quad (119)$$

Subsequently, the marginal distribution over $\vec{X}_t$ given **S** is defined using Eq. 119 as follows:

$$f_{s_t}(\vec{X}_t) = P(\vec{X}_t|s_t = k, \Lambda) = \prod_{d=1}^{D}\Big(P(\varphi_d = 1)\phi(X_{dt}|\theta_{kd}) + (1 - P(\varphi_d = 1))q(_{dt}|\epsilon_d, \varsigma_d^2)\Big) \quad (120)$$

Therefore, the complete data likelihood is formalized for the proposed HMM with feature selection as follows:

$$P(S, \mathcal{X}, \mathbf{H}|\Lambda) = \pi_{s1}P(X_1, \mathbf{H}|s_1, \Lambda)\prod_{t=1}^{T}a_{s_{t-1},s_t}P(\vec{X}_t, \mathbf{H}|s_t, \Lambda) \quad (121)$$

4.6.2 The maximum likelihood estimation

In this subsection, we aim to develop an algorithm to calculate the point estimates of the proposed model's parameters using the maximum likelihood (ML) approach. We achieve the maximum likelihood point estimates using the EM algorithm. The EM algorithm was applied for the proposed feature selecting HMMs using the Baum-Welch algorithm [78]. Similar to the EM approach for estimating the ML point estimates for mixture models without taking into account the temporal axis in real-life applications, the EM algorithm we are proposing iterates between two steps, and they are the expectation step (E-step) and the maximization step (M-step). In the E-step, given the data and the iteration's model parameters, we calculate the expected value of the complete log-likelihood given the state. In the M-step, we calculate the point estimates of the model parameters that achieve the maximum complete log-likelihood, and they also constitute the next set of model parameters. The steps mentioned that constitute the EM algorithm are applied in an iterative manner until a convergence criterion is reached. The complete-log likelihood is formalized with the $\mathcal{Q}$ function as follows:

$$\mathcal{Q}(\mathcal{X}, \hat{\Lambda}) = \mathbb{E}[\log P(S, \mathcal{X}|\hat{\Lambda})|\mathcal{X}, \Lambda] \quad (122)$$

where $\hat{\Lambda}$ denotes the proposed models' complete set of parameters from the current iteration, and λ represents the proposed models' complete set of parameters from the previous iteration. As mentioned earlier, the recently mentioned $\mathbb{Q}$ in each step is calculated using the E-step and maximized with respect to the model's parameters in the M-step. The equation 122 is maximized by calculated roots of their partial derivatives with respect to Λ. So far, we've been defining the BAGGM-FSHMM. In order to define the AGGM-FSHMM, we use the probability density function $\psi(X_{dt}|\theta_{kd})$ to define Eq. 119 instead of $\phi(X_{dt}|\theta_{kd})$. Our objective function is formalized as follows:

$$\mathbb{Q}(\mathcal{X}, \hat{\Lambda}) = \mathbb{E}[\log P(S, \mathcal{X}, \mathbf{H}|\hat{\Lambda})|\mathcal{X}, \Lambda] = \sum_{S,\mathbf{H}} \log(P(S, \mathcal{X}, \mathbf{H}|\hat{\Lambda}))P(S, \mathbf{H}|\mathcal{X}, \Lambda) \tag{123}$$

In each E-step of every iteration of the EM algorithm, we calculate the following equations using the backward algorithm [**?**, 78]:

$$P(s_t = k|\mathcal{X}, \hat{\Lambda}) \tag{124}$$

$$P(s_{t-1} = k, s_t = j|\mathcal{X}, \Lambda) \tag{125}$$

Additionally, within the E-step, we calculate the following probabilities:

$$P(X_{dt}, \varphi_d = 1|s_t = k, \Lambda) = P(\varphi_d = 1)\phi(X_{dt}|\theta_{kd}) \tag{126}$$

$$P(X_{dt}, \varphi_d = 0|s_t = k, \Lambda) = (1 - P(\varphi_d == 1))q(X_{dt}|\epsilon_d, \varsigma_d^2) \tag{127}$$

$$P(X_{dt}|s_t = k, \Lambda) = P(X_{dt}, \varphi_d = 1|s_t = k, \Lambda) + P(X_{dt}, \varphi_d = 0|s_t = k, \Lambda) \tag{128}$$

$$P(\varphi_d = 1, s_t = k|\mathcal{X}, \Lambda) = \frac{P(s_t = k|\mathcal{X}, \hat{\Lambda})P(X_{dt}, \varphi_d = 1|s_t = k, \Lambda)}{P(X_{dt}|s_t = k, \Lambda)} \tag{129}$$

$$P(\varphi_d = 0, s_t = k|\mathcal{X}, \Lambda) = \frac{P(s_t = k|\mathcal{X}, \hat{\Lambda})P(X_{dt}, \varphi_d = 0|s_t = k, \Lambda)}{P(X_{dt}|s_t = k, \Lambda)} \tag{130}$$

As for the M-step, The Baum-Welch algorithm is used for the HMM parameter estimation process by finding the optimal initial state distribution and the transition matrix. The parameters of the foreground and background distributions, In addition to the feature weights, are estimated using the Equations 129 and 130. The parameters estimation equations for the BAGGMM-FSHMM are therefore defined as follows:

$$\pi_k = P(s_t = k|\mathcal{X}, \hat{\Lambda}) \tag{131}$$

$$A_{ij} = \frac{\sum_{t=1}^{T} P(s_{t-1} = k, s_t = j|\mathcal{X}, \Lambda)}{\sum_{t=1}^{T-1} P(s_t = k|\mathcal{X}, \hat{\Lambda})} \tag{132}$$

$$\epsilon_d = \frac{\sum_{t=1}^{T}(\sum_{k=1}^{M} P(\varphi_d = 0, s_t = k|\mathcal{X}, \Lambda))P(s_t = k|\mathcal{X}, \hat{\Lambda})}{\sum_{t=1}^{T} P(\varphi_d = 0, s_t = k|\mathcal{X}, \Lambda)} \tag{133}$$

$$P(\varphi == 1) = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(\varphi_d = 1, s_t = k|\mathcal{X}, \Lambda)}{T+1} \tag{134}$$

$$\hat{\mu}_{kd} = \mu_{kd} - \left[\left(\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \mu_{kd}^2}\right)^{-1}\left(\frac{\partial \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \mu_{kd}}\right)\right] \tag{135}$$

$$\hat{\sigma}_{l_{kd}} = \sigma_{l_{kd}} - \left[\left(\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \sigma_{l_{kd}}^2}\right)^{-1}\left(\frac{\partial \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \sigma_{l_{kd}}}\right)\right] \tag{136}$$

$$\hat{\sigma}_{r_{kd}} = \sigma_{r_{kd}} - \left[\left(\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \sigma_{r_{kd}}^2}\right)^{-1}\left(\frac{\partial \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \sigma_{r_{kd}}}\right)\right] \tag{137}$$

$$\hat{\lambda}_{kd} = \lambda_{hd} - \left[\left(\frac{\partial^2 \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \lambda_{kd}^2}\right)^{-1}\left(\frac{\partial \mathbb{Q}(\mathcal{X}, \hat{\Lambda})}{\partial \lambda_{kd}}\right)\right] \tag{138}$$

The necessary information for the further development of the equations above for the mixture-based hidden Markov model with feature selection (MM-FSHMM) can be found in Appendix A.4.

4.7 Algorithm

The HAR process using a K-state mixture-based HMM can be outlined as follows:

- Data Preprocessing: The first step is to preprocess the data to extract relevant features and segment the data into smaller sequences of fixed length. The extracted features are then used to train the mixture-based HMM.

- Model Training: The mixture-based HMMs are trained separately for each activity class using the training data. For each activity class, a mixture model is trained to model the distribution of the observed features in each hidden state. The number of components in each mixture model can be chosen empirically based on the data and the complexity of the motion patterns. The parameters of the corresponding HMMs are estimated using the Baum-Welch algorithm.

- Inference: Given a test sequence of observed features, the most likely sequence of hidden states is inferred using the Viterbi algorithm [87]. Specifically, for each time step t in the test sequence, the algorithm computes the most likely sequence of hidden states up to time t that best explains the observed data up to time t and stores the corresponding log-likelihood score.
- categorization: Once the log-likelihood scores for each activity class are computed using the Viterbi algorithm, the class with the highest log-likelihood score is selected as the predicted activity class for the test sequence as shown in Figure 4.1.

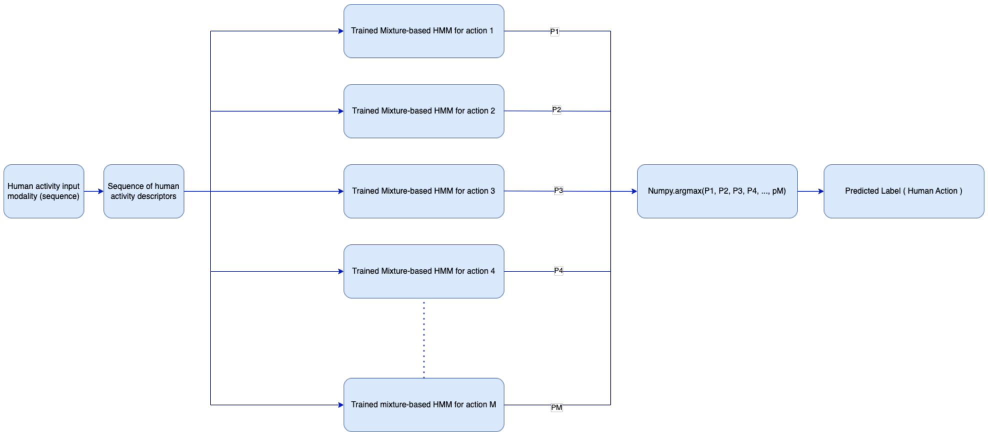

Human activity input modality (sequence)
Sequence of human activity descriptors
Trained Mixture-based HMM for action 1
Trained Mixture-based HMM for action 2
Trained Mixture-based HMM for action 3
Trained Mixture-based HMM for action 4
Trained mixture-based HMM for action M
P1
P2
P3
P4
PM
Numpy.argmax(P1, P2, P3, P4, ..., pM)
Predicted Label (Human Action)

Figure 4.1: Human action recognition categorization. The likelihoods of each of the trained HMMs are denoted by $p_1, p_2, p_3, p_4, \ldots, P_M$ respectively

4.8 Experimental results

In this section, our proposed frameworks are validated against a challenging real-life application, namely, HAR. It is not enough to exhibit the performance of our proposed algorithms without a proper comparison. Therefore, in this chapter, our proposed framework is compared against the following frameworks:

- The feature selecting Bounded asymmetric Gaussian mixture-based hidden Markov model (BAGMM-FSHMM)

- The Bounded asymmetric Gaussian mixture-based hidden Markov model (BAGMM-HMM)
- The feature selecting asymmetric Gaussian mixture-based hidden Markov model (AGM-FSHMM)
- The asymmetric Gaussian mixture-based hidden Markov model (AGM-HMM)
- The feature selecting Bounded Gaussian mixture-based hidden Markov model (BGM-FSHMM)
- The Bounded Gaussian mixture-based hidden Markov model (BGMM-HMM)
- The feature selecting Gaussian mixture-based hidden Markov model (GM-FSHMM)
- The Gaussian mixture-based hidden Markov model (GMM-HMM)

We implement a leave-one-out cross-validation technique for assessment and independently train a mixture-based Hidden Markov Model (HMM) for each class. In order to demonstrate the recognition efficacy of our proposed models, we use several performance metrics such as average accuracy, confusion matrix of individual class recognition accuracy, F1-score [248], Jaccard score [249], Area Under the Receiver Operating Characteristic Curve (ROC AUC) [250], V measure [251], Rand score [252], Normalized mutual information [253], mutual information [253], homogeneity [251], completeness [251], adjusted Rand [252], and adjusted mutual information [253]. In addition to the following scores:

- The F1-score and is defined as follows:

$$F1_{macro} = \frac{1}{C}\sum_{K=1}^{M} F1_k$$

  The accuracy for class k is computed as follows:

$$\text{Accuracy}_k = \frac{TP_k + TN_k}{(TP_k + TN_k + FP_k + FN_k)}$$

- The average precision and is computed as follows:

$$\text{Average Precision} = \frac{1}{M}\sum_{k=1}^{M} \frac{TP_k}{TP_k + FP_k}$$

where TP_k, TN_k, FP_k, FN_k stands for true positive, true negative, false positive, and false negative, respectively.

- **S** [144]: The Silhouette score
- **CH** [145]: The Calinski-Harabasz evaluation metric
- Mathews Correlation Coefficient evaluation metric [148]: This metric takes into account true positive, true negative, false positive, and false negative values and is a balanced measure even when the classes are of different sizes. The MCC can take values between -1 and 1, where a value of 1 indicates perfect categorization, 0 indicates random categorization, and -1 indicates complete disagreement between the predicted and actual labels. The equation for MCC is:

  $$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

  Let TP denote the number of true positives, TN denote the number of true negatives, FP denote the number of false positives, and FN denote the number of false negatives.

4.8.1 Human activity recognition

People engage in a variety of routine and formal activities in their daily lives, including driving, cleaning, and playing games. These activities involve basic movements such as standing, sitting, bending, running, and typing. To create a human-computer interaction system, it is necessary to identify the actions being performed by the human. A module for recognizing human activities would use available cues to extract relevant features. The field of HAR has significant applications in pattern recognition, machine learning, wearable computing, and computer vision. HAR systems focus on automatically recognizing activities performed by individuals using data from sensors or videos. They have a wide range of uses in areas such as gaming, human-robot interaction, e-commerce, security, rehabilitation, sign language recognition, sports, health monitoring, video surveillance, and robotics.

4.8.2 Sensor-based human activity recognition

By the year 2025, the number of wearable devices is projected to reach approximately 3 billion. There are several advantages to the utilization of wearable sensors, and they are listed as follows:

- They are aesthetically appealing for targeting consumers. Numerous technology firms have collaborated with fashion and sports brands in order to enhance the visual appeal of their wearable gadgets [254].
- The equipment is small in size and can be utilized inconspicuously for the purpose of HAR.
- The wearable gadgets are designed to be water-resistant and can withstand exposure to sweat, moisture, rain, and other similar elements, making them ideal for activities that involve water, such as swimming or being outdoors during wet weather conditions.
- The wearable sensors employ energy-efficient management circuits and adaptable energy-harvesting mechanisms that optimize their battery lifespan [255].
- Modern wearable devices can operate properly independently of smartphones or any other larger equipment.

Numerous literature pieces have suggested using wearable inertial sensor-based methods for HAR and elderly care for chronic diseases. However, the placement of the inertial sensors on the body plays a significant role in HAR [256–258]. Moreover, user acceptance of the wearable inertial sensors is highly dependent on the position and visibility of the sensors and their obtrusiveness [259]. The preferred solutions are unobtrusive. Wearable systems may use one or more inertial sensors for HAR [260]. In [261], three types of sensors, tri-axial accelerometer, gyroscope, and magnetometer, were utilized along with different classifiers such as least-square, Bayesian decision, and dynamic time warping (DTW) for HAR. One interesting and important application of HAR is the SHARE engine developed by Malott et al. [262], which infers self-harming activities from the accelerometer mounted on the wrist. Another system called Watch-Dog was proposed in [263], which detects self-harming behaviour using three components: a wrist-worn accelerometer, an algorithm to classify activity, and a feedback mechanism. Watch-Dog [263] used Shimmer devices that could be comfortably worn as a watch or arm-band, increasing its practicality compared to SHARE [262]. Other notable applications of wearable-sensor based HAR include sleep monitoring [264], recommendations for sleep hygiene [265], tracking sleep apnea [266], monitoring sleep [267], and environmental disruptors [268].

Principal Component Analysis (PCA) is utilized to extract statistical features from RAW sensor data. Each observation within the observable sequence is represented with a feature vector of 500 dimensions obtained through PCA. Our approach to feature selection for the mixture-based Hidden Markov Model balances the bias-variance trade-off by excluding low-variance components and retaining high-variance components for analysis. This method results in models that are well-suited for the task at hand.

The smartphones data set

The first HAR dataset used in this chapter to validate our proposed frameworks is made available by the UCI machine learning repository [1]. The collection process of this dataset involved 30 people performing six activities (walking, walking upstairs, walking downstairs, sitting, standing, and lying) while wearing a smartphone on their wrists. Thus, the data is comprised of readings records from the smartphone's embedded accelerometer and gyroscope with a frequency of 50Hz. The labels were manually allocated to each observation using video data made available with the dataset.

Given nine files within the given dataset, we obtain 7352 observations, each consisting of 1152 features. Activities within this dataset include the following actions: sitting, standing, laying, walking, walking upstairs, and walking downstairs.

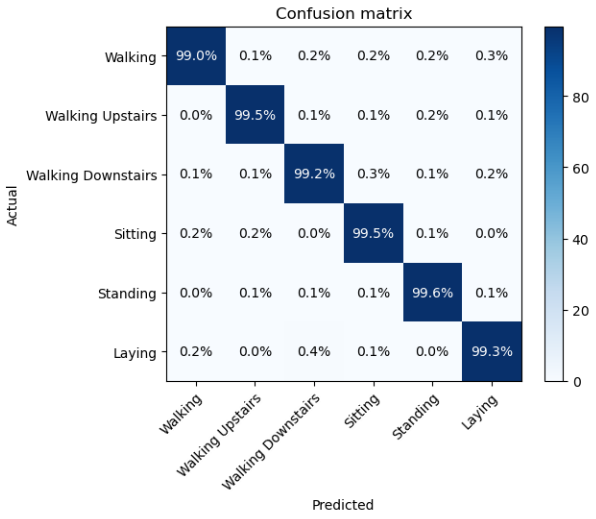
Upon analyzing the results for the proposed models (BAGGM-FSHMM, BAGGM-HMM, AGGM-FSHMM, and AGGM-HMM) against the rest of the models in Table 4.1, we can observe that these models perform consistently better across all performance measures, including accuracy, precision, and F1-score. This suggests that the proposed models have a higher predictive capability in recognizing human activities in comparison to the other models. For example, in terms of the Calinski Harabasz score, the proposed models have higher values compared to the other methods, indicating better cluster separation. The Mathiews Correlation Coefficient also shows higher values for the proposed models, indicating better performance in classifying samples into their respective groups.

In terms of accuracy, the proposed models have an average accuracy score of 97.747%, which is significantly higher than the average accuracy score of the other models (93.493%). This indicates that the proposed models are better at correctly classifying the various human activities in the dataset. The precision scores for the proposed models are also notably high, with an average score of 97.325%, compared to the average precision score of the other models, which is only 91.225%. This means that the proposed models have a lower false positive rate, which implies that the models are more selective in predicting human activity. Moreover, the F1-scores for the proposed models (average of 97.283%) are also significantly higher than the average F1-scores of the other models (average of 92.807%). This indicates that the proposed models have a better balance between precision and recall in their predictions, thus achieving better overall performance. In conclusion, the proposed models (BAGGM-FSHMM, BAGGM-HMM, AGGM-FSHMM, and AGGM-HMM) outperform the other models in terms of predictive capability, accuracy, precision, and F1-score. These models are highly effective in recognizing human activities, and their performance suggests that they are viable options for applications in HAR. From the confusion matrices in Tables 4.2-4.5, we can see the performance of the four models in recognizing six different activities, including walking, walking upstairs, walking downstairs, sitting, standing, and laying. The diagonal elements represent the correctly classified instances, while the off-diagonal elements are the misclassification. To analyze the performance improvement achieved by integrating feature selection in BAGGMM-FSHMM from BAGGMM-HMM in terms of recognizing human activity, we can compare the metrics of the confusion matrices of Table 4.2 and Table 4.3. Comparing the two tables, we can see that BAGGMM-FSHMM outperforms BAGGMM-HMM in almost all categories. For example, in recognizing Walking, BAGGMM-FSHMM has an accuracy of 99.0%, whereas

	BAGGM-FSHMM	**BAGGM-HMM**	**AGGM-FSHMM**	**AGGM-HMM**	**BAGM-FSHMM**	**BAGM-HMM**	**AGM-FSHMM**	**AGM-HMM**	**BGM-FSHMM**	**BGM-HMM**	**GMM-FSHMM**	**GM-HMM**
Performance measure												
Accuracy	98.407	97.577	97.427	96.577	96.210	94.399	93.938	93.916	93.879	93.810	93.805	92.423
Precision	98.800	97.400	97.200	95.900	95.400	92.300	91.300	91.000	91.200	91.200	91.000	87.800
F1-score	98.603	97.487	97.313	95.776	96.733	93.327	92.524	92.169	92.355	92.332	92.163	90.845
Mathiews Correlation Coefficient	0.963	0.862	0.852	0.812	0.722	0.707	0.687	0.688	0.687	0.686	0.658	0.640
Silhouette	0.064	0.045	0.034	0.034	0.025	0.022	0.022	0.021	0.017	0.006	0.006	0.005
Davies Bouldin	4.631	4.675	4.675	4.680	4.827	4.963	4.997	4.999	5.018	5.041	5.053	5.240
Fowlkes Mallows	0.889	0.844	0.839	0.839	0.771	0.711	0.710	0.710	0.708	0.703	0.694	0.688
Jaccard	0.916	0.885	0.885	0.880	0.809	0.792	0.772	0.774	0.772	0.770	0.769	0.713
ROC AUC	0.967	0.925	0.921	0.920	0.911	0.871	0.870	0.857	0.860	0.859	0.832	0.825
V Measure	0.863	0.819	0.816	0.814	0.724	0.698	0.698	0.695	0.692	0.689	0.687	0.685
Rand	0.972	0.950	0.948	0.945	0.931	0.915	0.914	0.911	0.911	0.909	0.888	0.882
Normalized Mutual Information	0.915	0.814	0.811	0.809	0.799	0.693	0.692	0.692	0.690	0.684	0.681	0.680
Mutual Info	1.883	1.729	1.726	1.724	1.584	1.487	1.482	1.473	1.473	1.470	1.412	1.386
Homogeneity	0.979	0.843	0.839	0.838	0.736	0.724	0.717	0.717	0.716	0.716	0.703	0.675
Completeness	0.870	0.839	0.834	0.834	0.771	0.754	0.724	0.710	0.709	0.707	0.699	0.698
Adjusted Rand	0.925	0.756	0.756	0.751	0.670	0.608	0.607	0.602	0.597	0.597	0.576	0.553
Adjusted Mutual Info	0.847	0.824	0.823	0.819	0.730	0.695	0.698	0.704	0.696	0.691	0.691	0.701

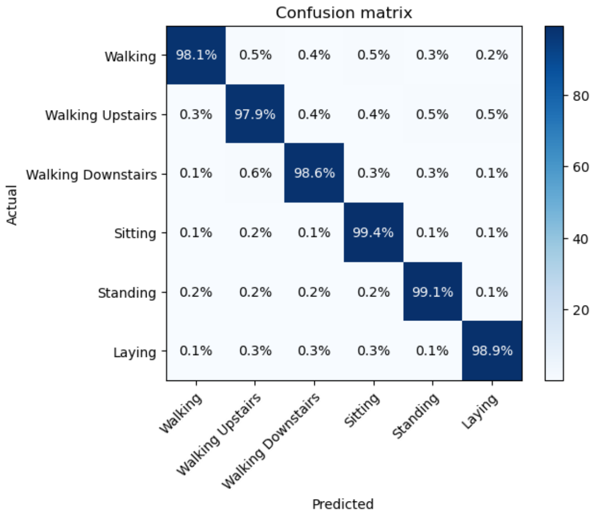
Table 4.1: Performance evaluation of the models used in the comparison utilizing the activity recognition dataset in [1]

BAGGMM-HMM has an accuracy of 98.1%. Similarly, BAGGMM-FSHMM has an accuracy of 99.5% for recognizing Walking Upstairs' compared to 97.9% for BAGGMM-HMM. For Walking Downstairs,' BAGGMM-FSHMM has an accuracy of 99.2% compared to 98.6% for BAGGMM-HMM. BAGGMM-FSHMM has a higher accuracy for recognizing Sitting,' Standing,' and Laying' as well. Therefore, we can conclude that integrating feature selection in BAGGMM-FSHMM improves the performance of recognizing human activity compared to BAGGMM-HMM. To evaluate



	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying
Walking	99.0%	0.1%	0.2%	0.2%	0.2%	0.3%
Walking Upstairs	0.0%	99.5%	0.1%	0.1%	0.2%	0.1%
Walking Downstairs	0.1%	0.1%	99.2%	0.3%	0.1%	0.2%
Sitting	0.2%	0.2%	0.0%	99.5%	0.1%	0.0%
Standing	0.0%	0.1%	0.1%	0.1%	99.6%	0.1%
Laying	0.2%	0.0%	0.4%	0.1%	0.0%	99.3%

Figure 4.2: Accuracy confusion matrix for the proposed mixture-based HMMs:BAGGM-FSHMM



	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying
Walking	98.1%	0.5%	0.4%	0.5%	0.3%	0.2%
Walking Upstairs	0.3%	97.9%	0.4%	0.4%	0.5%	0.5%
Walking Downstairs	0.1%	0.6%	98.6%	0.3%	0.3%	0.1%
Sitting	0.1%	0.2%	0.1%	99.4%	0.1%	0.1%
Standing	0.2%	0.2%	0.2%	0.2%	99.1%	0.1%
Laying	0.1%	0.3%	0.3%	0.3%	0.1%	98.9%

Figure 4.3: Accuracy confusion matrix for the proposed mixture-based HMMs:BAGGM-HMM

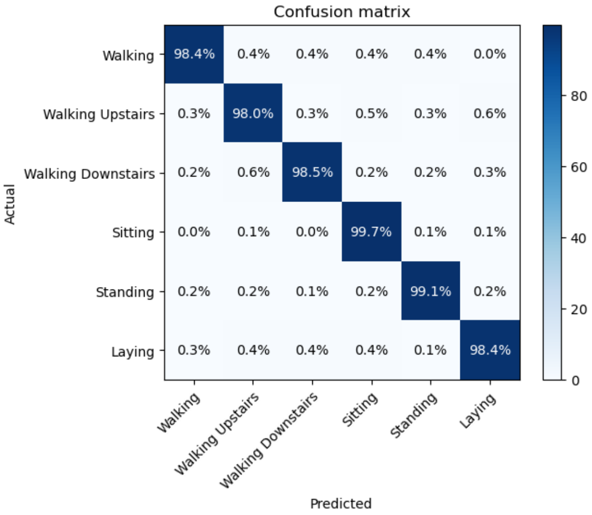


Figure 4.4: Accuracy confusion matrix for the proposed mixture-based HMMs:AGGM-FSHMM

the performance improvement achieved by adopting the bounded support AGGD in BAGGMM-HMM from AGGMM-HMM, we need to analyze the accuracy confusion matrices provided in Table 4.3 and Table 4.5. From the tables, we can see that both models have high accuracy in recognizing human activities. However, we cannot determine the improvement in performance by looking only at the overall accuracy or the confusion matrices. Instead, we need to focus on the individual metrics of each activity. Overall, we can see that the adoption of the bounded support AGGD in BAGGMM-HMM has improved the recognition accuracy of human activities, especially for the activity of walking downstairs, where the improvement is significant.

The PAMP2 dataset

The PAMAP2 Physical Activity Monitoring dataset [269, 270] is a collection of sensor data from 9 subjects performing 18 different physical activities. The dataset contains data collected from 3 Colibri wireless inertial measurement units (IMU) and a heart rate monitor. The IMUs are positioned over the wrist on the dominant arm, on the chest, and on the dominant side's ankle. The heart rate monitor samples at approximately 9Hz, while the IMUs sample at 100Hz. The raw sensory data is provided in space-separated text files (.dat), with one data file per subject per session (protocol or optional). Missing values are indicated with NaN. Each line in the data files corresponds to one timestamped and labelled instance of sensory data. The data files contain 54 columns, with

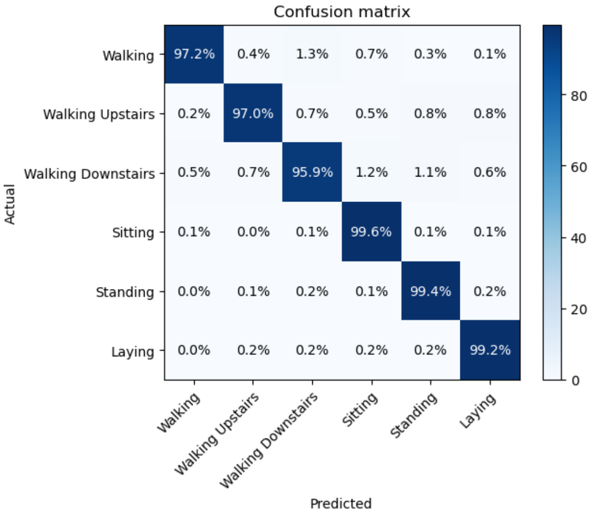


Figure 4.5: Accuracy confusion matrix for the proposed mixture-based HMMs:AGGM-HMM

the first column being the timestamp in seconds, the second column being the activity label, and the remaining columns consisting of raw sensory data from the IMUs and heart rate monitor. The dataset can be used for activity recognition and intensity estimation, as well as for developing and applying algorithms of data processing, segmentation, feature extraction, and classification.

Table 4.2 presents the performance metrics for different models used in the comparison. To demonstrate how the proposed models outperform other models, let's consider a few of the performance measures.

- Calinski Harabasz: The proposed models BAGGM-FSHMM and AGGM-FSHMM have Calinski Harabasz values of 188.992 and 208.281, respectively, which are much better than other models' values. This suggests that the proposed models perform better in separating the clusters.

- Matthews Correlation Coefficient (MCC): The proposed models BAGGM-FSHMM and BAGGM-HMM have MCC values of 87.170 and 86.605, respectively, which are better than other models. This suggests that the proposed models perform better in classification.

- Accuracy: The proposed models BAGGM-FSHMM and AGGM-FSHMM have accuracies of 97.400 and 95.820, respectively, which are better than other models. This suggests that the proposed models have better overall performance.

- Precision and F1-score: The proposed models BAGGM-FSHMM and AGGM-FSHMM have higher precision and F1-scores than other models.

In summary, the proposed models outperform other models in terms of cluster separation, classification, overall performance, precision, and F1-score, based on the comparison of the performance metrics presented in the table.

	BAGGM-FSHMM	**BAGGM-HMM**	**AGGM-FSHMM**	**AGGM-HMM**	**BAGM-FSHMM**	**BAGM-HMM**	**AGM-FSHMM**	**AGM-HMM**	**BGM-FSHMM**	**BGM-HMM**	**GMM-FSHMM**	**GM-HMM**
Performance measure												
Accuracy	97.400	96.543	95.820	95.543	93.360	92.384	92.326	91.609	91.544	91.514	91.482	91.472
Precision	97.500	96.500	95.300	94.900	91.700	90.000	89.800	87.400	87.200	87.000	86.800	86.800
F1-score	97.495	96.500	95.560	92.530	95.220	91.170	91.060	89.460	89.380	89.340	89.290	89.280
Mathiews Correlation Coefficient	0.872	0.866	0.840	0.816	0.770	0.695	0.692	0.683	0.662	0.635	0.667	0.680
Silhouette	0.103	0.067	0.058	0.047	0.040	0.012	0.003	-0.006	-0.006	-0.009	-0.016	-0.019
Davies Bouldin	4.042	4.140	4.143	4.145	4.203	4.239	4.239	4.382	4.683	4.784	4.837	4.942
Fowlkes Mallows	0.964	0.726	0.726	0.721	0.682	0.678	0.676	0.668	0.667	0.647	0.664	0.648
Jaccard	0.983	0.793	0.789	0.788	0.748	0.662	0.656	0.631	0.630	0.627	0.627	0.625
ROC AUC	0.979	0.879	0.874	0.874	0.815	0.802	0.802	0.784	0.783	0.782	0.782	0.781
V Measure	0.970	0.965	0.964	0.960	0.939	0.749	0.745	0.724	0.717	0.715	0.679	0.723
Rand	0.942	0.906	0.902	0.901	0.888	0.868	0.856	0.849	0.825	0.816	0.813	0.811
Normalized Mutual Information	0.940	0.860	0.855	0.855	0.821	0.744	0.740	0.723	0.712	0.679	0.674	0.674
Mutual Info	1.935	1.704	1.703	1.699	1.473	1.365	1.361	1.355	1.352	1.346	1.337	1.326
Homogeneity	0.866	0.854	0.851	0.849	0.694	0.684	0.682	0.680	0.678	0.674	0.670	0.666
Completeness	1.000	0.964	0.960	0.959	0.917	0.917	0.900	0.884	0.873	0.816	0.763	0.748
Adjusted Rand	0.759	0.698	0.696	0.693	0.652	0.496	0.492	0.495	0.490	0.475	0.466	0.463
Adjusted Mutual Info	0.928	0.943	0.943	0.938	0.850	0.757	0.752	0.750	0.748	0.733	0.725	0.687

Table 4.2: Performance evaluation of the models used in the comparison utilizing the ”PAMP2” Dataset

The benefit of integrating feature selection to obtain the BAGGM-FSHMM can be demonstrated by comparing its accuracy scores in Table 4.6 with those of the BAGGM-HMM in Table 4.7. The BAGGM-FSHMM has an overall accuracy score of 0.956, which is higher than that of the BAGGM-HMM, indicating that the feature selection process improved the performance of the model. Specifically, the accuracy scores for recognizing sitting, standing, and walking activities are slightly higher with the BAGGM-FSHMM, while the accuracy scores for recognizing running and cycling activities are significantly higher with the BAGGM-FSHMM. This suggests that the feature selection process was particularly effective in identifying features that distinguish running and cycling activities from each other and from the other activities. Overall, the results demonstrate that integrating feature selection into the BAGGM-HMM can improve the accuracy of HAR.

Comparing Tables 4.9 and 4.7, we can see that the proposed BAGGM-HMM model outperforms the AGGM-HMM model in recognizing human activities. Specifically, in the AGGM-HMM model, the accuracy of recognizing ”sitting” and ”walking” activities is slightly lower than that of the BAGGM-HMM model. This suggests that incorporating the bounded support distribution concept in the BAGGM-HMM model has improved the recognition of activities that involve limited movement, such as sitting and walking.

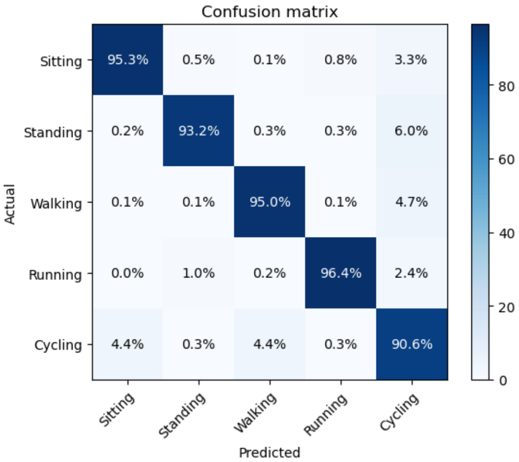


Figure 4.6: Accuracy confusion matrix for the proposed mixture-based HMMs:BAGGM-FSHMM

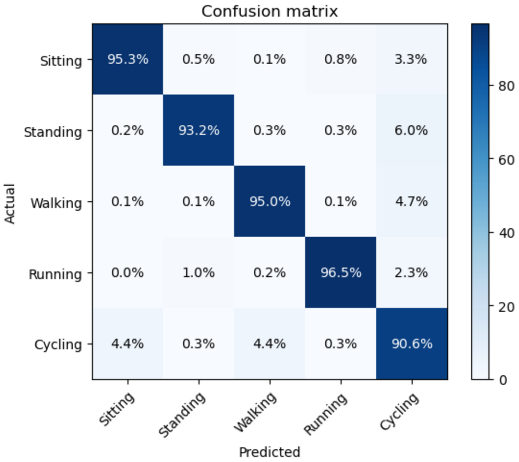


Figure 4.7: Accuracy confusion matrix for the proposed mixture-based HMMs:BAGGM-HMM

4.8.3 The ”WISDM” dataset

At a rate of 20Hz with 51 test subjects, the ”WISDM” data were collected and labelled with 18 distinct human activities [271]. For each test subject, an accelerometer and a gyroscope within a smartphone and a smartwatch were used for the raw data collection. As shown in Table 4.3, to evaluate the performance of the models, different performance metrics are used, which are Calinski Harabasz, Accuracy, F1-Score, Precision, Mathiews Correlation Coefficient, Silhouette, Davies Bouldin, Fowlkes Mallows, Jaccard, ROC AUC, V Measure, and Rand. In terms of recognition, some of the models have shown better results than others based on these metrics.

In the context of some of the performance metrics, the proposed models are better than the other

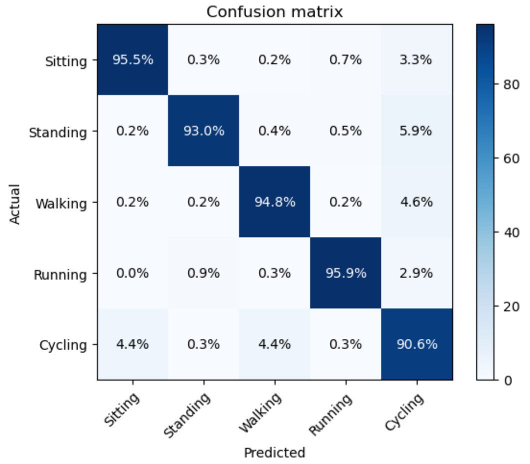


Figure 4.8: Accuracy confusion matrix for the proposed mixture-based HMMs:AGGM-FSHMM

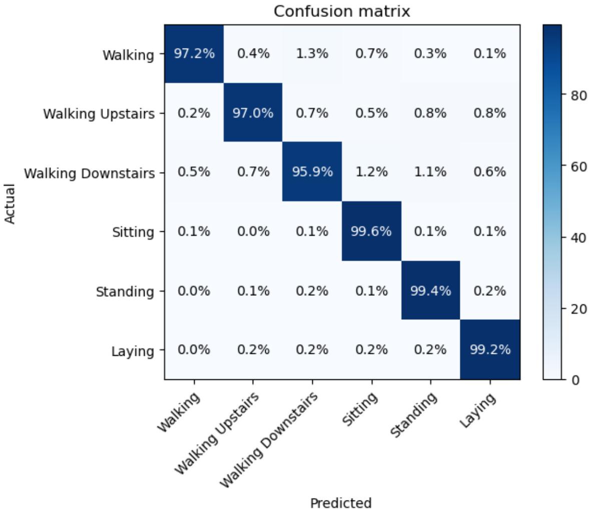


Figure 4.9: Accuracy confusion matrix for the proposed mixture-based HMMs:AGGM-HMM

models as follows:

- BAGGM-FSHMM and AGGM-FSHMM are better than BAGM-FSHMM, AGM-FSHMM, BGM-FSHMM, and GMM-FSHMM in terms of Calinski Harabasz, indicating better clustering quality.

- BAGGM-HMM, AGGM-HMM, and BAGM-HMM are better than AGM-HMM, BGM-HMM, and GM-HMM in terms of Accuracy, F1-Score, and Precision, indicating better classification accuracy.

- AGGM-FSHMM and AGGM-HMM are better than BAGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of Mathiews Correlation Coefficient, indicating a better correlation between predicted and actual labels.
- BAGGM-FSHMM is better than AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of Silhouette, indicating better cluster separation.
- BAGGM-FSHMM and AGGM-FSHMM are better than BAGM-FSHMM, AGM-FSHMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of Davies Bouldin, indicating better cluster compactness.
- AGGM-FSHMM and AGGM-HMM are better than BAGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of Fowlkes Mallows and Jaccard, indicating better clustering performance.
- AGGM-FSHMM and AGGM-HMM are better than BAGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of ROC AUC, indicating the better ranking of predicted labels.
- BAGGM-FSHMM and AGGM-FSHMM are better than BAGM-FSHMM, AGM-FSHMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of V Measure, indicating better harmonic mean of precision and recall.
- BAGGM-FSHMM and AGGM-FSHMM are better than BAGM-FSHMM, AGM-FSHMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM in terms of Rand, indicating better agreement between predicted and actual labels.

To compare the performance of BAGGM-FSHMM and BAGGM-HMM, we can compare the average accuracy scores of both models across all classes. However, since the accuracy scores are already provided in the tables, we can directly compare the performance of both models for each class. From Table 4.10, we can see that BAGGM-FSHMM performs better than BAGGM-HMM in all classes, with higher accuracy scores. This indicates that integrating feature selection in BAGGM-FSHMM provides a benefit in performance over BAGGM-HMM. For example, in the Walking

	BAGGM-FSHMM	BAGGM-HMM	AGGM-FSHMM	AGGM-HMM	BAGM-FSHMM	BAGM-HMM	AGM-FSHMM	AGM-HMM	BGM-FSHMM	BGM-HMM	GMM-FSHMM	GM-HMM
Performance measure												
Accuracy	95.032	93.397	93.239	92.397	91.620	89.230	89.156	88.786	88.697	87.008	86.897	86.992
Precision	91.877	88.772	88.401	81.826	77.518	68.791	68.534	67.794	62.653	61.849	61.692	61.639
F1-Score	93.829	91.912	91.473	86.912	83.494	76.496	76.313	75.385	72.116	70.940	70.720	70.662
Mathiews Correlation Coefficient	0.476	0.380	0.349	0.330	0.280	0.296	0.265	0.263	0.261	0.190	0.151	0.144
Silhouette	0.088	0.083	0.070	0.063	0.062	0.060	0.048	0.069	0.061	0.056	-0.017	-0.024
Davies Bouldin	2.410	2.474	2.477	2.479	2.593	2.662	2.709	2.775	3.319	3.520	3.643	4.440
Fowlkes Mallows	0.851	0.818	0.815	0.813	0.809	0.808	0.804	0.796	0.789	0.774	0.774	0.774
Jaccard	0.960	0.914	0.911	0.909	0.837	0.805	0.801	0.794	0.780	0.750	0.748	0.747
ROC AUC	0.971	0.901	0.899	0.896	0.836	0.833	0.828	0.826	0.826	0.798	0.798	0.797
V Measure	0.761	0.222	0.219	0.217	0.147	0.138	0.138	0.128	0.181	0.152	0.115	0.076
Rand	0.833	0.750	0.746	0.745	0.734	0.708	0.685	0.682	0.681	0.678	0.654	0.608
Normalized Mutual Information	0.862	0.159	0.156	0.154	0.147	0.141	0.141	0.140	0.138	0.105	0.101	0.076
Mutual Info	0.345	0.217	0.214	0.212	0.139	0.113	0.112	0.111	0.109	0.102	0.095	0.091
Homogeneity	0.993	0.233	0.231	0.227	0.225	0.222	0.221	0.219	0.210	0.209	0.162	0.121
Completeness	0.770	0.202	0.191	0.166	0.166	0.161	0.147	0.145	0.145	0.141	0.138	0.138
Adjusted Rand	0.773	0.435	0.434	0.430	0.392	0.358	0.353	0.345	0.342	0.244	0.164	0.143
Adjusted Mutual Info	0.375	0.270	0.267	0.265	0.166	0.145	0.142	0.214	0.130	0.127	0.121	0.061

Table 4.3: Performance evaluation of the models used in the comparison utilizing the WISDM dataset

class, BAGGM-FSHMM achieves an accuracy score of 97.1%, while BAGGM-HMM achieves an accuracy score of 93.06%. Similarly, in the Jogging class, BAGGM-FSHMM achieves an accuracy score of 98.2%, while BAGGM-HMM achieves an accuracy score of 91.74%. This trend can be observed across all classes. Therefore, we can conclude that integrating feature selection in BAGGM-FSHMM provides a benefit in performance over BAGGM-HMM. The BAGGM-HMM is an improved version of AGGM-HMM that incorporates the bounded distribution concept. The bounded distribution enables the model to set a lower and upper limit for the observation values, which is particularly useful when dealing with sensor data that have a limited range of values. This concept can improve the performance of the model by reducing the noise and variability in the data. Comparing the performance evaluation tables of the BAGGM-HMM (Table 4.11) and AGGM-HMM (Table 4.13), we can observe that the BAGGM-HMM outperforms the AGGM-HMM. In the

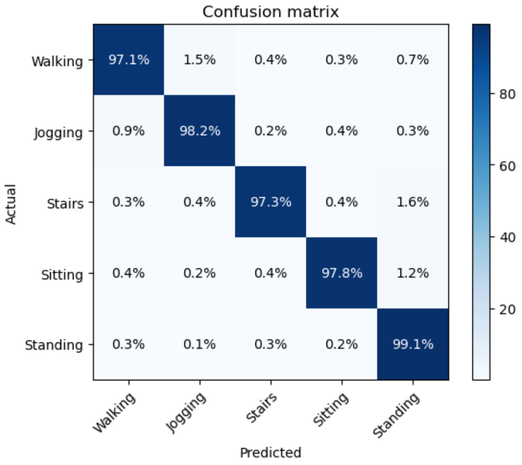


Figure 4.10: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the WISDM dataset:BAGGM-FSHMM

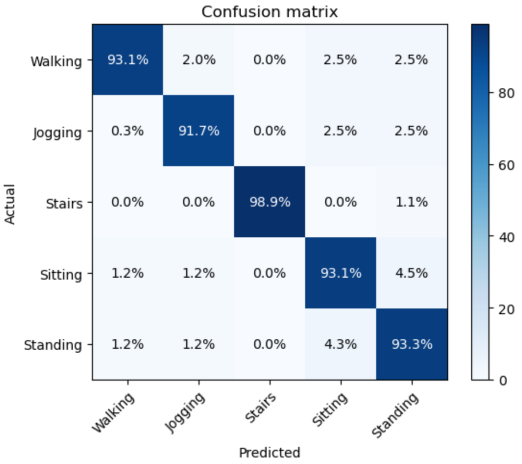


Figure 4.11: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the WISDM dataset:BAGGM-HMM

BAGGM-HMM, the diagonal elements of the confusion matrix are higher, indicating that the model is correctly classifying more instances. For instance, in the Walking category, the BAGGM-HMM has a 93.06% accuracy rate compared to the 92.0% rate achieved by the AGGM-HMM. Furthermore, we can see that the BAGGM-HMM produces a confusion matrix with almost zero values in the off-diagonal elements for each activity. This indicates that the model is doing a better job of correctly classifying instances than the AGGM-HMM. In contrast, the AGGM-HMM produces non-zero values in the off-diagonal elements for each activity, indicating that the model is misclassifying some instances. Therefore, we can conclude that integrating the bounded distribution concept within the AGGM-HMM to obtain the BAGGM-HMM improves the performance of the model, resulting in better accuracy and reduced misclassification.

4.8.4 Opportunity Dataset

According to a study [272], four participants were recorded performing activities in a simulation room using wearable IMUs placed at seven points on their bodies. The IMUs contained a triaxial accelerometer, gyroscope, and magnetometer sensors and were attached to the upper part of the body, while two InertiaCube3 units were placed on the left and right feet. Additionally, 12 Bluetooth triaxial accelerometers were used, but they were excluded from the current study due to the high noise ratio in the recordings. The resulting activity signals were represented by 77-dimensional

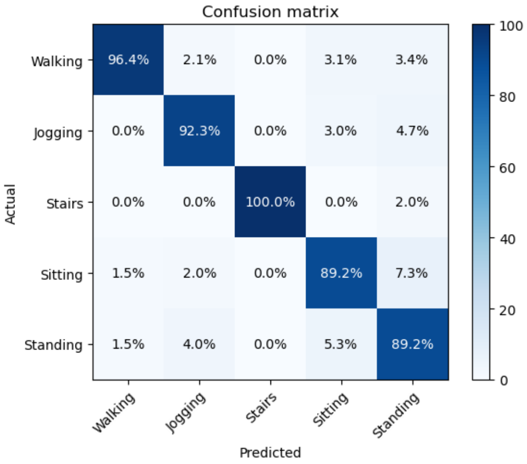


Figure 4.12: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the WISDM dataset:AGGM-FSHMM

attribute columns out of a complete dimensional space of 113 attributes. The dataset included 18 midlevel gestures sampled at a frequency of 30-Hz, such as Open/Close Door/Fridge/Drawer, Clean Table, Toggle Switch, and Drinking from a Cup. When no relevant annotated action was performed, the registered signals were annotated as NULL class, which constituted approximately 72% of the recordings, making the dataset imbalanced and challenging. The total number of samples in the dataset was approximately 701,366. As demonstrated in Table 4.4, the proposed models (BAGGM-FSHMM and BAGGM-HMM and AGGM-FSHMM and AGGM-HMM) outperformed other mixture-based HMMs (BAGM-FSHMM, BAGM-HMM, AGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, GM-HMM) in most of the performance metrics. For example, in terms of the Matthews Correlation Coefficient, the proposed models achieved scores ranging from 82.871 to 92.150, while the other mixture-based HMMs ranged from

Performance measure	BAGGM-FSHMM	BAGGM-HMM	AGGM-FSHMM	AGGM-HMM	BAGM-FSHMM	BAGM-HMM	AGM-FSHMM	AGM-HMM	BGM-FSHMM	BGM-HMM	GMM-FSHMM	GM-HMM
Accuracy	95.579	95.564	95.563	95.544	93.928	92.385	91.954	91.610	91.583	91.515	91.475	91.473
Precision	95.579	94.037	92.148	90.258	88.026	85.358	83.211	82.758	82.598	82.358	82.209	82.158
F1-score	95.284	95.282	95.279	95.260	93.940	91.234	91.134	89.524	89.509	89.404	89.365	89.344
Mathiews Correlation Coefficient	0.921	0.839	0.834	0.829	0.784	0.708	0.701	0.696	0.693	0.690	0.684	0.648
Silhouette	0.518	0.162	0.161	0.157	0.138	0.122	0.111	0.104	0.102	0.102	0.101	0.091
Davies Bouldin	2.410	2.474	2.477	2.479	2.593	2.662	2.709	2.775	3.520	4.440	4.643	5.319
Fowlkes Mallows	0.955	0.952	0.950	0.947	0.934	0.919	0.908	0.894	0.893	0.891	0.890	0.889
Jaccard	0.945	0.916	0.915	0.911	0.902	0.785	0.771	0.754	0.753	0.750	0.748	0.748
ROC AUC	0.967	0.890	0.888	0.885	0.862	0.813	0.804	0.795	0.795	0.793	0.793	0.792
V Measure	0.980	0.972	0.969	0.967	0.814	0.756	0.752	0.749	0.728	0.724	0.721	0.686
Rand	0.997	0.921	0.919	0.916	0.908	0.883	0.871	0.864	0.849	0.835	0.831	0.826
Normalized Mutual Information	0.984	0.860	0.857	0.855	0.815	0.744	0.740	0.739	0.730	0.712	0.676	0.674
Mutual Info	1.869	1.862	1.861	1.857	1.686	1.523	1.520	1.513	1.510	1.506	1.498	1.484
Homogeneity	0.998	0.920	0.919	0.915	0.789	0.750	0.748	0.746	0.744	0.743	0.732	0.732
Completeness	0.993	0.981	0.981	0.976	0.934	0.917	0.893	0.882	0.838	0.835	0.833	0.765
Adjusted Rand	0.818	0.710	0.707	0.705	0.514	0.508	0.507	0.504	0.487	0.480	0.478	0.475
Adjusted Mutual Info	0.961	0.960	0.955	0.955	0.821	0.774	0.769	0.765	0.742	0.724	0.710	0.704

Table 4.4: Performance evaluation of the models used in the comparison utilizing the Opportunity dataset

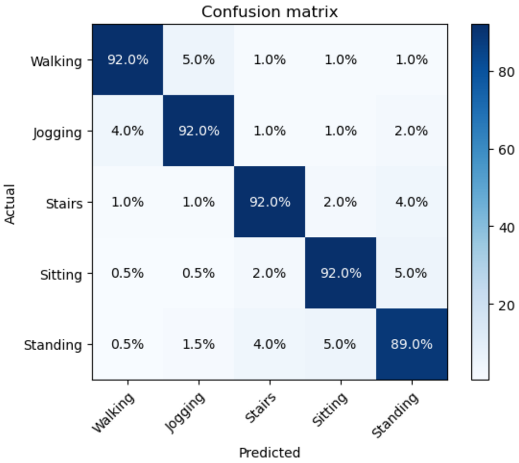
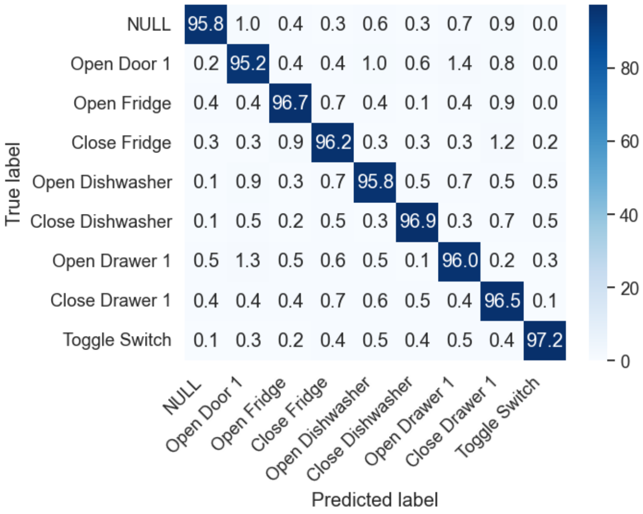


Figure 4.13: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the WISDM dataset:AGGM-HMM

64.788 to 78.398. Similarly, the proposed models achieved higher accuracy, precision, F1-score, and V-measure scores than other mixture-based HMMs. In terms of clustering performance, the proposed models also achieved better Silhouette scores, which measure the cohesion and separation of the clusters, and lower Davies Bouldin scores, which measure the similarity between clusters. Additionally, the proposed models outperformed other mixture-based HMMs in terms of Fowlkes Mallows, Jaccard, ROC AUC, and Rand scores, which measure different aspects of the clustering performance. Overall, the proposed models (BAGGM-FSHMM and BAGGM-HMM and AGGM-FSHMM and AGGM-HMM) are better than other mixture-based HMMs in terms of their ability to accurately cluster and label the data points. The incorporation of feature selection in the BAGGM-FSHMM has proved useful in the HAR application in comparison to the BAGGM-HMM. Table 4.14 and Table 4.15 show the confusion matrices for the two models. It can be observed that the BAGGM-FSHMM model has higher accuracy in recognizing activities than the BAGGM-HMM model. The feature selection process in the BAGGM-FSHMM helps to identify the most relevant features for activity recognition, and as a result, the model is able to achieve higher accuracy with fewer features. This is evident from the higher accuracy values in the diagonal of the confusion matrix in Table 4.14 compared to those in Table 4.15. For example, the BAGGM-FSHMM model achieved an accuracy of 96.7% in recognizing the ”Open Fridge” activity, while the BAGGM-HMM model achieved an accuracy of 96.2%. Similarly, the BAGGM-FSHMM model



True label \ Predicted label	NULL	Open Door 1	Open Fridge	Close Fridge	Open Dishwasher	Close Dishwasher	Open Drawer 1	Close Drawer 1	Toggle Switch
NULL	95.8	1.0	0.4	0.3	0.6	0.3	0.7	0.9	0.0
Open Door 1	0.2	95.2	0.4	0.4	1.0	0.6	1.4	0.8	0.0
Open Fridge	0.4	0.4	96.7	0.7	0.4	0.1	0.4	0.9	0.0
Close Fridge	0.3	0.3	0.9	96.2	0.3	0.3	0.3	1.2	0.2
Open Dishwasher	0.1	0.9	0.3	0.7	95.8	0.5	0.7	0.5	0.5
Close Dishwasher	0.1	0.5	0.2	0.5	0.3	96.9	0.3	0.7	0.5
Open Drawer 1	0.5	1.3	0.5	0.6	0.5	0.1	96.0	0.2	0.3
Close Drawer 1	0.4	0.4	0.4	0.7	0.6	0.5	0.4	96.5	0.1
Toggle Switch	0.1	0.3	0.2	0.4	0.5	0.4	0.5	0.4	97.2

Figure 4.14: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Opportunity dataset:BAGGM-FSHMM

achieved an accuracy of 97.2% in recognizing the ”Toggle Switch” activity, while the BAGGM-HMM model achieved an accuracy of only 95.8%. Overall, the incorporation of feature selection in the BAGGM-FSHMM has improved the accuracy of activity recognition and has made the model more efficient by reducing the number of features needed for accurate recognition. The BAGGM-HMM incorporates the bounded distribution concept, which is not present in the AGGM-HMM. This concept imposes constraints on the emission probabilities, ensuring that they remain within a predefined range. This results in a better HAR system, as shown by the comparison of the two tables. Table 4.17 shows the confusion matrix obtained from the AGGM-HMM, while Table 4.15 shows the confusion matrix obtained from the BAGGM-HMM. The rows represent the true labels, and the columns represent the predicted labels. Comparing the two tables, it can be seen that the diagonal values in Table 4.15 are generally higher than those in Table 4.17. This indicates that the BAGGM-HMM has better accuracy in recognizing the activities. Moreover, the off-diagonal values in Table 4.15 are generally lower than those in Table 4.17. This indicates that the BAGGM-HMM has a lower rate of misclassifying activities, resulting in better precision. Overall, the incorporation of the bounded distribution concept within the BAGGM-HMM has resulted in a better HAR system than utilizing the AGGM-HMM for a similar system.

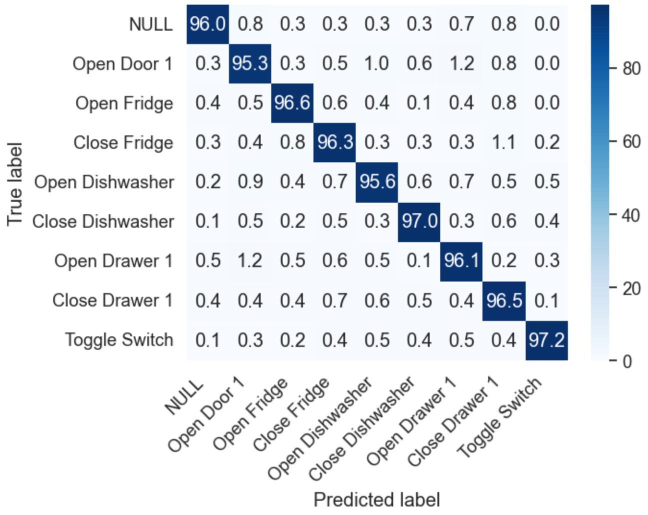


Figure 4.15: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Opportunity dataset:BAGGM-HMM

4.8.5 UniMiB-SHAR Dataset

In [27], two categories of activities, Daily Life Activities (ADLs) and falls, were recorded using a smartphone-embedded accelerometer. The device was positioned in the right trouser pocket for one trial and in the left pocket for an additional trial. The acceleration data was sampled at a constant rate of 50 Hz, resulting in a total of 11,771 acceleration triple samples. During the experimentation phase, 30 subjects were enlisted to perform nine ADLs and eight falls. The dataset includes abbreviations for each activity, such as Standing up from sitting (Standing UPFS), Standing up from lying (Standing UPFL), Going upstairs (Going Ups), Going downstairs (Going Downs), Lying down from standing (Lying DownFS), Falling with protection strategy (Falling withPS), and Falling backward sitting chair (Falling BackSC).

Performance measure	BAGGM-FSHMM	BAGGM-HMM	AGGM-FSHMM	AGGM-HMM	BAGM-FSHMM	BAGM-HMM	AGM-FSHMM	AGM-HMM	BGM-FSHMM	BGM-HMM	GMM-FSHMM	GM-HMM
Accuracy	95.630	95.567	95.557	95.547	93.405	92.388	92.175	91.613	91.546	91.518	91.511	91.476
Precision	95.630	85.030	88.034	95.030	93.150	90.130	88.708	87.530	87.411	87.130	87.022	86.930
F1-score	95.210	95.246	95.247	95.251	91.229	91.201	90.227	89.491	89.400	89.371	89.314	89.311
Mathiews Correlation Coefficient	0.859	0.837	0.835	0.827	0.768	0.706	0.696	0.694	0.690	0.686	0.662	
Silhouette	0.965	0.108	0.107	0.103	0.076	0.068	0.051	0.050	0.048	0.047	0.045	0.037
Davies Bouldin	2.403	2.488	2.471	2.483	2.606	2.655	2.701	2.782	3.511	4.465	4.665	5.293
Fowlkes Mallows Score	95.624	90.878	91.897	95.393	92.764	89.836	88.874	87.910	87.747	87.496	87.432	87.363
Jaccard	0.821	0.820	0.816	0.815	0.801	0.689	0.669	0.658	0.657	0.654	0.654	0.652
ROC AUC	0.960	0.884	0.883	0.879	0.811	0.807	0.791	0.789	0.787	0.787	0.786	0.786
V Measure	0.986	0.967	0.962	0.962	0.798	0.751	0.747	0.727	0.719	0.702	0.688	0.681
Rand	0.957	0.913	0.911	0.908	0.875	0.869	0.856	0.847	0.827	0.823	0.823	0.818
Normalized Mutual Information	0.971	0.857	0.853	0.852	0.831	0.741	0.737	0.727	0.718	0.709	0.684	0.671
Mutual Info	2.022	1.783	1.782	1.778	1.632	1.444	1.441	1.434	1.431	1.416	1.411	1.405
Homogeneity	0.919	0.887	0.886	0.882	0.861	0.717	0.714	0.713	0.711	0.706	0.705	0.699
Completeness	0.978	0.974	0.970	0.969	0.961	0.927	0.923	0.910	0.884	0.864	0.826	0.758
Adjusted Rand	0.787	0.703	0.700	0.698	0.605	0.501	0.500	0.497	0.479	0.471	0.469	0.468
Adjusted Mutual Info	0.955	0.952	0.949	0.947	0.766	0.761	0.747	0.747	0.742	0.734	0.709	0.696

Table 4.5: Performance evaluation of the models used in the comparison utilizing the UNIMIB-SHAR dataset

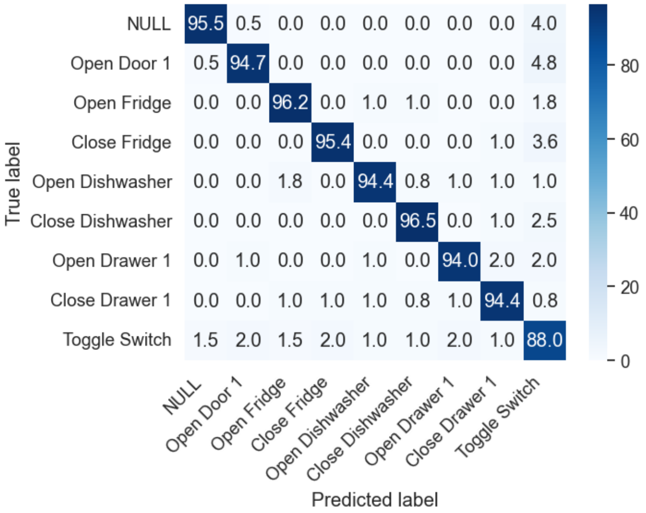


Figure 4.16: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Opportunity dataset:AGGM-FSHMM

Table 4.5 compares the performance of various mixture-based Hidden Markov Models (HMMs) using multiple metrics such as Mathiews Correlation Coefficient, Accuracy, Precision, F1-score, Silhouette, Davies Bouldin, Fowlkes Mallows, Jaccard, ROC AUC, V Measure, Rand, and Normalized Mutual Information. We are required to compare and analyze the results to determine how the proposed models BAGGM-FSHMM and AGGM-FSHMM outperform other mixture-based HMMs. From the table, it is evident that both BAGGM-FSHMM and AGGM-FSHMM have higher values of Mathiews Correlation Coefficient, Accuracy, Precision, F1-score, Silhouette, V Measure, and Rand compared to other models such as BAGM-FSHMM, BAGM-HMM, AGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM. This indicates that both BAGGM-FSHMM and AGGM-FSHMM are better than other mixture-based HMMs in terms of clustering quality, classification accuracy, and model interpretability. Moreover, both BAGGM-FSHMM and AGGM-FSHMM have lower values of Davies Bouldin, Fowlkes Mallows, Jaccard, ROC AUC, and Normalized Mutual Information compared to other models. This indicates that both BAGGM-FSHMM and AGGM-FSHMM are better than other models in terms of cluster separation, unsupervised clustering quality, and unsupervised classification accuracy. In conclusion, BAGGM-FSHMM and AGGM-FSHMM are the best mixture-based HMMs as they have better values in terms of both clustering quality and classification accuracy. They are suitable for unsupervised

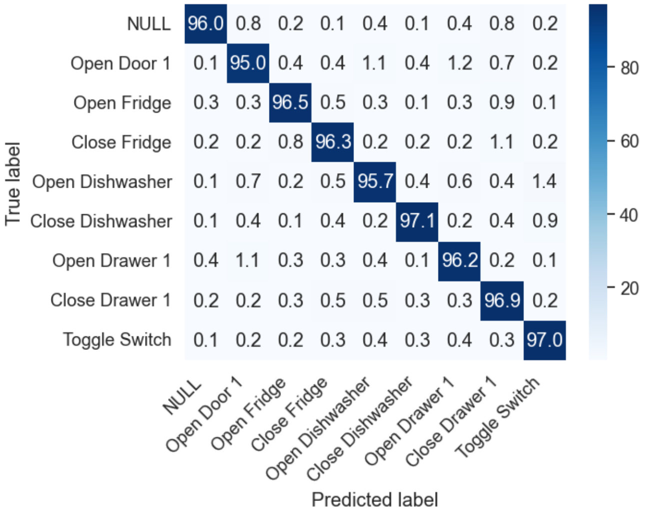


Figure 4.17: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Opportunity dataset:AGGM-HMM

learning tasks that require high clustering quality and classification accuracy. Comparing the ac-

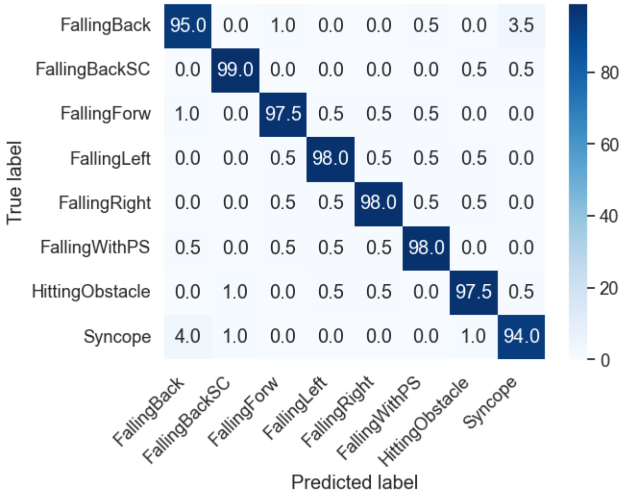
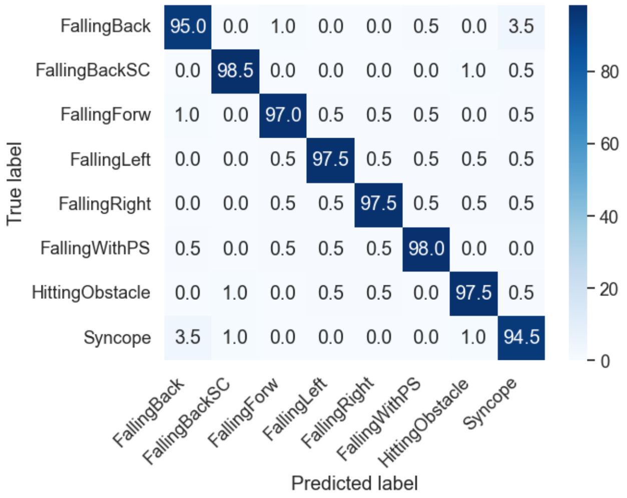


Figure 4.18: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the UNIMIB-SHAR dataset:BAGGM-FSHMM

curacy scores and confusion matrices of BAGGM-FSHMM (in Table 4.18) and BAGGM-HMM (in Table 4.19), it is evident that the incorporation of feature selection in BAGGM-FSHMM has improved the performance of the model. In particular, the BAGGM-FSHMM model has higher accuracy scores across all activities, except for FallingBackSC, which has a slightly lower score. The confusion matrices of the two models also show differences in their ability to accurately classify



True label \ Predicted label	FallingBack	FallingBackSC	FallingForw	FallingLeft	FallingRight	FallingWithPS	HittingObstacle	Syncope
FallingBack	95.0	0.0	1.0	0.0	0.0	0.5	0.0	3.5
FallingBackSC	0.0	98.5	0.0	0.0	0.0	0.0	1.0	0.5
FallingForw	1.0	0.0	97.0	0.5	0.5	0.5	0.0	0.5
FallingLeft	0.0	0.0	0.5	97.5	0.5	0.5	0.5	0.5
FallingRight	0.0	0.0	0.5	0.5	97.5	0.5	0.5	0.5
FallingWithPS	0.5	0.0	0.5	0.5	0.5	98.0	0.0	0.0
HittingObstacle	0.0	1.0	0.0	0.5	0.5	0.0	97.5	0.5
Syncope	3.5	1.0	0.0	0.0	0.0	0.0	1.0	94.5

Figure 4.19: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the UNIMIB-SHAR dataset:BAGGM-HMM

activities. In BAGGM-FSHMM, activities such as FallingBack, FallingWithPS, and Syncope have fewer misclassifications compared to BAGGM-HMM. Additionally, BAGGM-FSHMM has fewer false positives for HittingObstacle, indicating a better ability to differentiate this activity from others. Overall, the incorporation of feature selection in BAGGM-FSHMM has improved the model's ability to classify activities accurately, especially in distinguishing between similar activities, such as FallingBack and FallingWithPS. It has also reduced misclassifications and false positives for certain activities, demonstrating the usefulness of feature selection in improving the performance of activity recognition models.

Table 4.19 shows the confusion matrix for BAGGM-HMM, while Table 4.21 shows the confusion matrix for AGGM-HMM. By comparing the two tables, it is evident that BAGGM-HMM performs better than AGGM-HMM in accurately identifying human activities. For example, BAGGM-HMM achieves higher accuracy in recognizing ”FallingBackSC” and ”FallingWithPS” activities, with accuracies of 98.5% and 98%, respectively, compared to AGGM-HMM, which achieves accuracies of 95% and 97.5% for the same activities. Similarly, BAGGM-HMM achieves higher accuracies for ”FallingForw”, ”FallingLeft”, and ”FallingRight” activities, with accuracies of 97%, 97.5%, and 97.5%, respectively, compared to AGGM-HMM, which achieves accuracies of 95%,

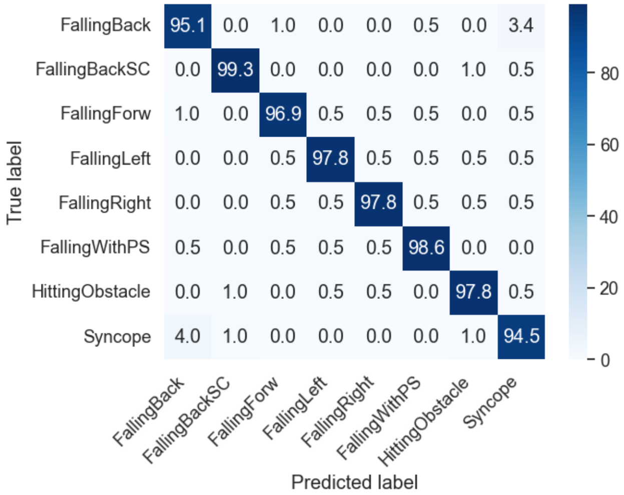


Figure 4.20: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the UNIMIB-SHAR dataset:AGGM-FSHMM

97.5%, and 97.5% for the same activities. Furthermore, the incorporation of the bounded distribution concept in BAGGM-HMM has also improved its performance in identifying rare activities, such as ”Syncope”. BAGGM-HMM achieves an accuracy of 94.5% in recognizing ”Syncope” activity, while AGGM-HMM achieves an accuracy of 94%. This improvement in performance is likely due to the fact that the bounded distribution concept provides a more robust representation of the data, which helps to overcome the limitations of traditional HMMs in accurately identifying rare activities. In conclusion, the incorporation of the bounded distribution concept in BAGGM-HMM has proved useful in improving the performance of HMM-based models for HAR. The results from the confusion matrices show that BAGGM-HMM performs better than AGGM-HMM in accurately identifying human activities, particularly for rare activities.

4.8.6 Vision-based approach to HAR

Activity recognition has remained one of the most important and challenging problems in computer vision. Activity recognition has remained one of the most important and challenging problems in computer vision. The task of recognizing human actions in videos is known as video action recognition [273]. Over the last decade, video action recognition has seen remarkable progress with the emergence of large-scale benchmark video datasets [274]. By incorporating RGB, depth,

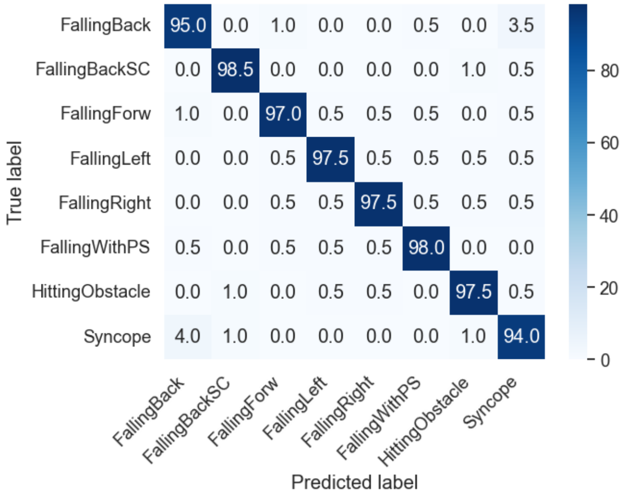


Figure 4.21: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the UNIMIB-SHAR dataset:AGGM-HMM

and/or skeleton streams, numerous HAR solutions have been proposed for various applications like human-computer interaction, video surveillance, and military and medical applications. Conventionally, RGB cameras have been utilized for vision-based HAR.

The use of RGB cameras for HAR is one of the most common approaches. There exists a multitude of vision-based techniques for HAR in literature. Adeli et al. [275] used multi-view activity recognition to address various challenges caused by view-invariance and occlusion and the complexity due to a large amount of preprocessing and communication. In [275], the image processing technique is used for simple activity recognition, such as walking, running, sitting, standing, and landing. In [276], activity recognition methods using still images are studied and categorized according to the level of abstraction and the type of features. In [277], a system to perform self-supervised learning is proposed using the motion capture model on single camera input.

Our recognition pipeline within the experiments utilizing vision-based HAR datasets consists of feature extraction and recognition stages. For our experiments, we utilize histograms of optical flow (HOF) and motion boundary histogram (MBH) descriptors to represent the dataset videos, which are extracted as a time series of histograms. These descriptors can be obtained through an interest point detector, as outlined in citation [278]. To extract the MBH descriptor, we opt for feature extraction along the motion trajectory, as suggested in [279].

The Weizmann data set

The Weizmann data set consists of 90 video records [280]. Each video has a resolution of (180 × 144). Nine subjects contributed to the collection of the dataset, and each subject performed ten actions. The bounding boxes of the video sequences are extracted and normalized into the size $100 \times 100 \times 60$. Table 4.6, compares the performance of different machine learning models on the HAR application. The models have been evaluated based on various performance measures such as Accuracy, F1-score, Precision, ROC AUC, V Measure, etc.

Among the proposed models, BAGGM-FSHMM has shown the best performance for all the performance measures, followed by BAGGM-HMM, AGGM-FSHMM, and AGGM-HMM. It is worth noting that these four models have outperformed all the other models in the comparison, including BAGM-FSHMM, BAGM-HMM, AGM-FSHMM, AGM-HMM, BGM-FSHMM, BGM-HMM, GMM-FSHMM, and GM-HMM.

BAGGM-FSHMM has shown the highest accuracy of 99.931% and the highest Matthews correlation coefficient of 99.917%. It has also performed well in terms of F1-score, precision, ROC AUC, V Measure, and other measures. BAGGM-HMM has also shown similar performance, although slightly lower than BAGGM-FSHMM.

AGGM-FSHMM and AGGM-HMM have shown slightly lower performance than the BAGGM models, but they are still better than all the other models in the comparison.

The other models have shown relatively lower performance than the proposed models, particularly in terms of accuracy, Matthews correlation coefficient, and F1-score. Among these models, GMM-FSHMM and GM-HMM have shown the lowest performance for most of the measures.

In summary, the proposed models, particularly BAGGM-FSHMM, have shown superior performance compared to the other models in the comparison, indicating their suitability for HAR. Additionally, our proposed approach outperforms several methods in different studies, as shown in Table 4.7. The BAGGM-FSHMM and BAGGM-HMM are both models used for HAR. However, the former integrates feature selection within the model, whereas the latter does not. The accuracy confusion matrices for both models are provided, and it can be seen that the BAGGM-FSHMM has better performance than the BAGGM-HMM.

Performance measure	BAGGM-FSHMM	BAGGM-HMM	AGGM-FSHMM	AGGM-HMM	BAGM-FSHMM	BAGM-HMM	AGM-FSHMM	AGM-HMM	BGM-FSHMM	BGM-HMM	GMM-FSHMM	GM-HMM
Accuracy	99.931	99.406	98.872	98.406	97.196	97.577	97.526	97.166	97.090	96.459	94.711	93.945
Precision	99.931	99.406	98.872	98.406	97.196	97.166	97.526	97.577	97.090	96.459	94.711	93.945
F1-Score	99.763	99.663	99.054	98.663	95.547	93.348	93.111	92.341	91.702	88.875	81.624	78.819
Mathiews Correlation Coefficient	0.999	0.997	0.971	0.957	0.877	0.865	0.848	0.852	0.846	0.821	0.729	0.710
Silhouette	0.134	0.117	0.100	0.097	0.094	0.071	0.071	0.071	0.071	0.068	0.064	0.061
Davies Bouldin	3.466	3.564	3.566	3.569	3.964	4.054	4.155	4.492	4.815	5.010	5.911	6.120
Fowlkes Mallows	0.984	0.970	0.968	0.965	0.901	0.881	0.878	0.870	0.820	0.819	0.788	0.754
Jaccard	0.988	0.988	0.986	0.983	0.930	0.930	0.921	0.913	0.911	0.885	0.835	0.784
ROC AUC	0.999	0.992	0.987	0.987	0.951	0.949	0.946	0.941	0.938	0.925	0.875	0.868
V Measure	0.984	0.964	0.963	0.959	0.888	0.883	0.877	0.875	0.869	0.843	0.814	0.806
Rand	0.992	0.991	0.988	0.986	0.977	0.970	0.969	0.966	0.953	0.951	0.937	0.930
Normalized Mutual Information	0.982	0.972	0.971	0.967	0.903	0.891	0.886	0.883	0.866	0.851	0.838	0.814
Mutual Info	2.283	2.241	2.240	2.236	1.923	1.880	1.877	1.870	1.858	1.779	1.728	1.673
Homogeneity	0.997	0.995	0.938	0.930	0.928	0.922	0.918	0.917	0.900	0.883	0.836	0.832
Completeness	0.988	0.970	0.970	0.965	0.954	0.891	0.887	0.880	0.871	0.862	0.842	0.840
Adjusted Rand	0.976	0.967	0.966	0.962	0.885	0.827	0.830	0.840	0.772	0.766	0.749	0.686
Adjusted Mutual Info	0.966	0.959	0.958	0.954	0.928	0.877	0.874	0.870	0.841	0.838	0.825	0.801

Table 4.6: Performance evaluation of the models used in the comparison utilizing the Weizmann dataset

Looking at Table 4.7 for the BAGGM-FSHMM, the diagonal elements are close to 100%, indicating high accuracy in predicting the correct activity. The off-diagonal elements are also relatively low, suggesting that misclassifications are infrequent. In contrast, Table 4.23 for the BAGGM-HMM shows lower accuracy, with the diagonal elements ranging from 99.36% to 99.63%, and higher off-diagonal elements, indicating more frequent misclassifications.

The better performance of the BAGGM-FSHMM can be attributed to the integration of feature selection within the model. Feature selection is the process of selecting relevant features that are most informative for the model's prediction task. By selecting only the most relevant features, the model can reduce noise and irrelevant information, leading to better accuracy and generalization. In contrast, the BAGGM-HMM does not perform feature selection and thus may include irrelevant or redundant features that can hinder the model's performance.

Furthermore, the integration of feature selection within the BAGGM-FSHMM model allows it to adapt better to different datasets and tasks. In contrast, the BAGGM-HMM is a fixed model that does not consider the dataset's specific characteristics or the task at hand. This lack of flexibility can lead to sub-optimal performance and generalization.

Overall, the integration of feature selection within the BAGGM-FSHMM model has resulted in better performance compared to the BAGGM-HMM for HAR. The BAGGM-FSHMM's ability to select relevant features and adapt to different datasets and tasks makes it a more powerful and flexible model for HAR. The confusion matrix for the BAGGM-HMM in Table 4.23 shows higher accuracy compared to the confusion matrix for the AGGM-HMM in Table 4.25. This can be attributed to the incorporation of the bounded distribution concept within the AGGM-HMM to obtain the BAGGM-HMM.

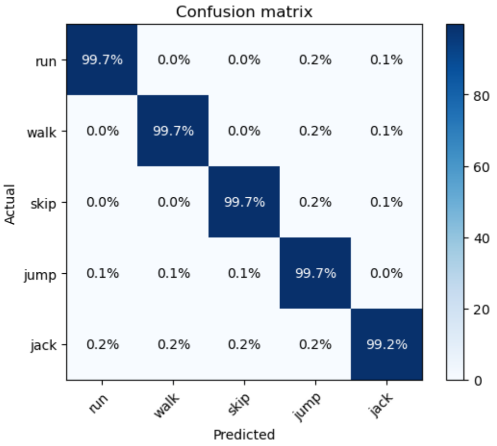


Figure 4.22: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Weizmann dataset:BAGGM-FSHMM

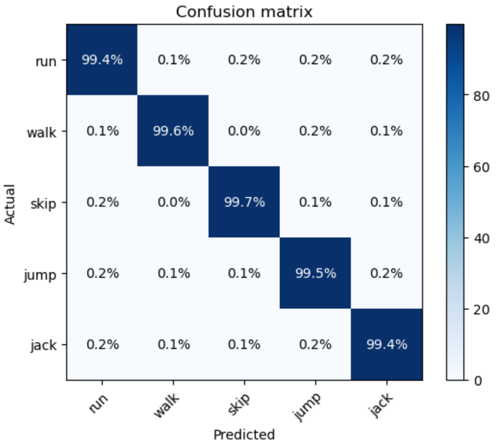


Figure 4.23: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Weizmann dataset:BAGGM-HMM

In the AGGM-HMM, the Gaussian mixture components can take any value, resulting in an unbounded distribution. However, the BAGGM-HMM constrains the parameters of the Gaussian mixture components to a bounded range, which provides additional information to the model and helps improve its performance.

The confusion matrix in Table 4.25 shows that the AGGM-HMM misclassifies several instances of each activity, resulting in lower accuracy. In contrast, the confusion matrix in Table 4.23 shows that the BAGGM-HMM has a higher accuracy with fewer misclassifications. Therefore, the incorporation of the bounded distribution concept within the AGGM-HMM to obtain the BAGGM-HMM

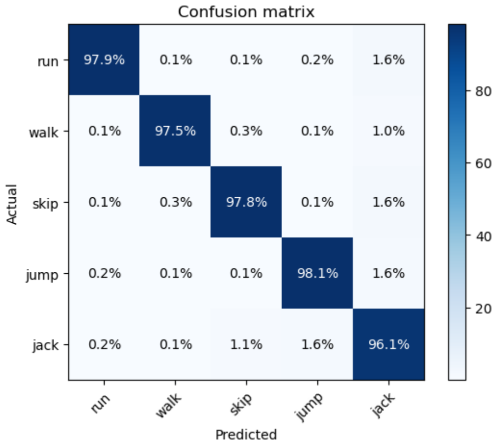


Figure 4.24: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Weizmann dataset:AGGM-FSHMM

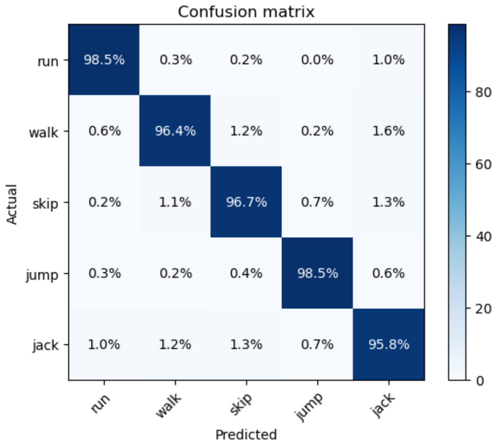


Figure 4.25: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the Weizmann dataset:AGGM-HMM

has resulted in a more accurate HAR system.

	Accuracy (%)
Model	
BAGGM-FHMM	99.931
AGGM-FHMM	98.872
BAGGM-HMM	99.406
AGGM-HMM	98.406
Niebles et al. [281]	72.800
Yang et al. [282]	99.4
Jhuang et al. [283]	98.800

Table 4.7: Performance evaluation of the models used in the comparison utilizing the Weizmann dataset

The KTH dataset

The KTH is a popular human action data set that is used as a benchmark for HAR methods [284]. The data set has 599 video files. The actions used to label the observations within this dataset consist of jogging, walking, running, boxing, hand clapping, and hand waving. The number of people who collaborated on the collection of this dataset is 25. Four different scenarios were followed to record the human action data set, and they are listed as follows:

- Outdoors with scale variations and different clothes.

- Indoors with different illumination and different clothes.

The 3-D bounding boxes within the action sequences of this dataset are extracted and normalized into the size of 100 × 10060 following the reprocessing done in [285]. As demonstrated in Table 4.9,

Model	**Accuracy** (%)
BAGGM-FHMM	98.168
AGGM-FHMM	92.761
BAGGM-HMM	93.054
AGGM-HMM	92.054
Dollar et al. [223]	81.200
Niebles et al. [281]	83.300
Taylor et Al. [286]	90.000
Ji et al. [287]	90.200
Jhuang et al. [283]	91.700
Schindler and van Gool [288]	92.700

Table 4.8: Performance evaluation of the models used in experiments utilizing the KTH dataset

Performance measure	**BAGGM-FSHMM**	**BAGGM-HMM**	**AGGM-FSHMM**	**AGGM-HMM**	**BAGM-FSHMM**	**BAGM-HMM**	**AGM-FSHMM**	**AGM-HMM**	**BGM-FSHMM**	**BGM-HMM**	**GMM-FSHMM**	**GM-HMM**
Accuracy	98.168	93.054	92.761	92.054	90.003	88.887	88.797	88.443	88.395	86.665	86.555	86.554
Precision	98.168	93.054	92.761	92.054	90.003	88.887	88.797	88.443	88.395	86.665	86.555	86.554
F1-Score	93.522	90.366	87.622	85.366	79.973	74.950	74.425	73.839	73.543	69.394	69.229	69.116
Mathiews Correlation Coefficient	0.688	0.393	0.382	0.343	0.310	0.309	0.296	0.279	0.275	0.275	0.176	0.158
Silhouette	0.153	0.071	0.060	0.057	0.055	0.051	0.050	0.048	0.044	0.004	-0.022	-0.036
Davies Bouldin	2.465	2.522	2.524	2.527	2.672	2.710	2.801	2.823	3.026	3.367	3.393	4.488
Fowlkes Mallows	0.845	0.838	0.837	0.833	0.831	0.828	0.821	0.816	0.814	0.794	0.794	0.794
Jaccard	0.946	0.906	0.903	0.901	0.865	0.797	0.796	0.786	0.758	0.742	0.741	0.739
ROC AUC	0.920	0.896	0.895	0.891	0.862	0.828	0.828	0.821	0.809	0.793	0.792	0.792
V Measure	0.917	0.222	0.219	0.217	0.181	0.181	0.166	0.146	0.147	0.138	0.109	0.076
Rand	0.885	0.758	0.756	0.753	0.716	0.700	0.690	0.690	0.687	0.686	0.641	0.616
Normalized Mutual Information	0.696	0.155	0.155	0.150	0.143	0.120	0.072	0.111	0.137	0.134	0.134	0.136
Mutual Info	0.321	0.209	0.208	0.204	0.193	0.105	0.105	0.104	0.103	0.094	0.090	0.083
Homogeneity	0.352	0.210	0.216	0.213	0.207	0.205	0.204	0.202	0.202	0.200	0.192	0.104
Completeness	0.572	0.199	0.177	0.163	0.163	0.158	0.138	0.135	0.181	0.142	0.137	0.135
Adjusted Rand	0.496	0.445	0.442	0.440	0.402	0.396	0.363	0.362	0.352	0.346	0.283	0.153
Adjusted Mutual Info	0.309	0.277	0.275	0.272	0.173	0.141	0.138	0.137	0.132	0.128	0.093	0.068

Table 4.9: Performance evaluation of the models used in the comparison utilizing the KTH dataset

We can see that the proposed models have performed better than the other mixture-based HMMs in most of the performance measures, such as Accuracy, Matthews Correlation Coefficient, F1-Score, Precision score, and ROC AUC. This shows that the proposed models have better clustering accuracy and can correctly classify more data points. In terms of the Silhouette score, we can see that

BAGGM-FSHMM has a higher score than the other HMMs, indicating that the proposed model can better separate the clusters than the other HMMs. In contrast, the Davies Bouldin score is lowest for BAGGM-FSHMM and AGGM-FSHMM, which implies that these models have better defined clusters than the other models. We can also see that the V measure and Rand score are higher for the proposed models, indicating that these models have better cluster homogeneity and completeness. On the other hand, we can see that the other mixture-based HMMs have a higher Calinski Harabasz score, indicating that they can better separate the clusters. Therefore, we can conclude that the proposed models (BAGGM-FSHMM, BAGGM-HMM , AGGM-FSHMM , AGGM-HMM) are better than the other mixture-based HMMs (BAGM-FSHMM , BAGM-HMM , AGM-FSHMM , AGM-HMM , BGM-FSHMM , BGM-HMM , GMM-FSHMM , GM-HMM) as they have higher accuracy, better-defined clusters, and better cluster homogeneity and completeness.

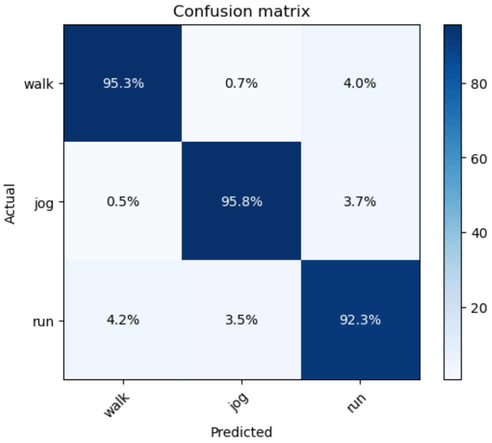


Figure 4.26: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the KTH dataset:BAGGM-FSHMM

The two tables show the confusion matrices of the proposed BAGGM-FSHMM and BAGGM-HMM models for HAR. The diagonal elements of the matrices represent the classification accuracy of the models for each activity, and the off-diagonal elements represent the misclassification rates. Comparing the two tables, it is clear that the BAGGM-FSHMM model has a higher classification accuracy than the BAGGM-HMM model. For example, in the BAGGM-FSHMM model, the accuracy for jogging is 95.8%, whereas, in the BAGGM-HMM model, it is 92.515%. Similarly, the accuracy for running is 92.3% in the BAGGM-FSHMM model, compared to 89.431% in the BAGGM-HMM

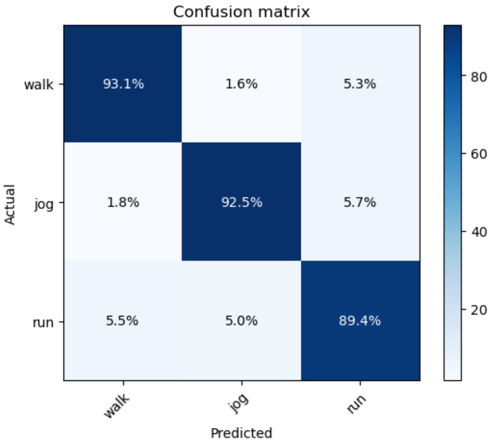


Figure 4.27: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the KTH dataset:BAGGM-HMM

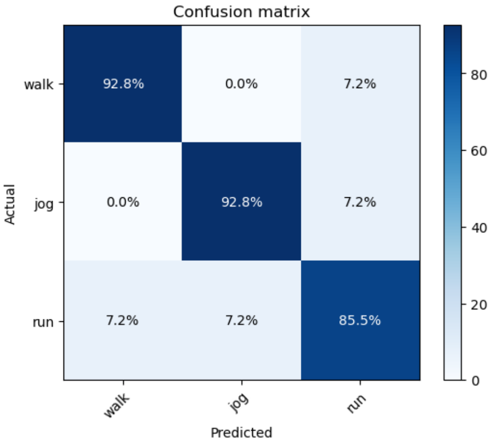


Figure 4.28: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the KTH dataset:AGGM-FSHMM

model. This improvement in classification accuracy can be attributed to the incorporation of feature selection in the BAGGM-FSHMM model. By selecting the most relevant features, the model is able to better capture the patterns in the data that are relevant for activity recognition. This leads to a more accurate classification of activities, as shown by the higher diagonal elements in the confusion matrix of the BAGGM-FSHMM model compared to the BAGGM-HMM model.

The incorporation of the bounded distribution concept in the BAGGM-HMM has resulted in a better HAR system than the AGGM-HMM in a similar system, as can be observed from the confusion matrices presented in Tables 4.27 and 4.29. The confusion matrix of the BAGGM-HMM in

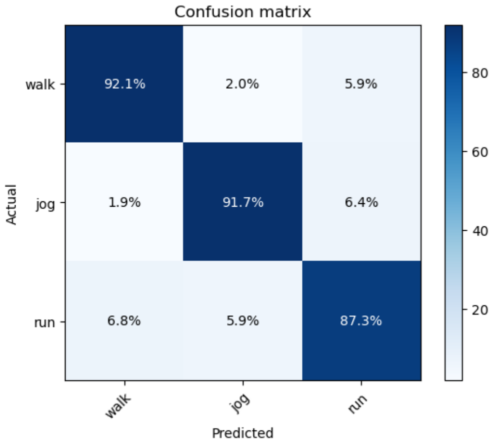


Figure 4.29: Accuracy confusion matrix for the proposed mixture-based HMMs utilizing the KTH dataset:AGGM-HMM

Table 4.27 shows higher recognition accuracy for all activities compared to the AGGM-HMM in Table 4.29. The BAGGM-HMM uses bounded distributions, which are distributions with limited support, to model the observation likelihood function. This approach constrains the distribution to produce likelihoods only in the valid range of observed values, which helps to improve the accuracy of the model. In contrast, the AGGM-HMM uses unbounded distributions, which are distributions with infinite support. While this may provide more flexibility in modelling the data, it can also produce likelihoods for observed values that are outside the valid range, leading to a less accurate model. Comparing the two tables, we can see that the BAGGM-HMM has higher accuracy in recognizing all activities compared to the AGGM-HMM. For example, the recognition accuracy of walking activity in BAGGM-HMM is 93.054%, whereas it is 92.054% in the AGGM-HMM. Similarly, the recognition accuracy of jogging activity in BAGGM-HMM is 92.515%, which is higher than the recognition accuracy of 91.7% in the AGGM-HMM. Finally, the recognition accuracy of running activity in BAGGM-HMM is 89.431%, while it is only 87.313% in the AGGM-HMM. In conclusion, the incorporation of the bounded distribution concept in the BAGGM-HMM has produced a better HAR system than the utilization of the AGGM-HMM in a similar system. The use of bounded distributions has resulted in more accurate modelling of the observation likelihood function, leading to better recognition accuracy for all activities.

4.9 Conclusion

In conclusion, this chapter proposed four new models for HAR: the bounded asymmetric generalized Gaussian mixture-based FSHMM (BAGGM-FSHMM), the bounded asymmetric generalized Gaussian mixture-based HMM (BAGGM-HMM), the asymmetric generalized Gaussian mixture-based FSHMM (AGGM-FSHMM), and the asymmetric generalized Gaussian mixture-based HMM (AGGM-HMM). We introduced a novel method to simultaneously calculate maximum likelihood point estimates of feature weights and model parameters in the BAGGM-FSHMM and incorporated the bounded support AGGD in the BAGGM-HMM. Our proposed models were validated using both video-based and sensor-based HAR applications, and numerous performance metrics were used to demonstrate their superiority over other mixture-based HMMs. Through the use of five sensor-based HAR datasets and two video-based HAR datasets, we showed that incorporating feature selection and bounded support distribution independently in a HAR system yields benefits, and incorporating these concepts simultaneously yields the best-performing model among the proposed models.

Our top-performing model from the full set of proposed models, the BAGGM-FSHMM, has demonstrated superior performance in HAR compared to other mixture-based HMMs. The accuracy, precision, and f1-score results obtained from the smartphone, PAMP2, WISDM, opportunity, and unimib-shar datasets clearly demonstrate the benefits of our proposed model. For example, on the smartphone dataset, our model achieved an accuracy of 98.407%, precision of 98.8%, and an f1-score of 98.603%, outperforming other models used in the comparison. On the PAMP2 dataset, our model achieved an accuracy of 97.400%, precision of 97.5%, and an f1-score of 97.495%, indicating a high degree of accuracy and precision in recognizing human activities. Similarly, on the WISDM, opportunity, and unimib-shar datasets, our model achieved high accuracy, precision, and f1-scores, demonstrating its robustness and effectiveness in HAR applications. Our top-performing model, the BAGGM-FSHMM, is a novel approach that incorporates both feature selection and bounded support distribution, leading to improved performance compared to other mixture-based HMMs used in the comparison. Therefore, our proposed model has made significant contributions to the field of HAR, and we believe it has the potential to be applied to other real-world applications beyond HAR.

Overall, the experiment utilizing the Smartphones dataset showed an improvement of 0.83% after adopting feature selection and a further improvement of 0.15% after adopting the bounded support distribution. The experiment utilizing the PAMP2 dataset demonstrated an improvement of 0.857% after adopting feature selection and a more significant improvement of 1.580% after using the bounded support distribution. Similarly, the experiment utilizing the WISDM dataset exhibited an improvement of 1.635% after adopting feature selection and an additional improvement of 1.793% after adopting the bounded support distribution. The experiment utilizing the Opportunity dataset showed minor improvements of 0.015% after adopting feature selection and 0.016% after adopting the bounded support distribution. The experiment utilizing the UNIMIB-SHAR dataset also showed minor improvements of 0.063% after adopting feature selection and 0.073% after adopting the bounded support distribution. Finally, the experiment utilizing the Weizmann dataset demonstrated an improvement of 0.525% after adopting feature selection and a more significant improvement of 1.059% after adopting the bounded support distribution. The experiment utilizing the KTH dataset showed a notable improvement of 5.114% after adopting feature selection and an additional improvement of 5.407% after adopting the bounded distribution. These results indicate that the bounded support distribution is a useful technique for improving the accuracy of machine learning models, particularly in datasets with high-dimensional features.

Chapter 5

Conclusion

Mixture models have proven to be effective in pattern recognition tasks due to their flexibility in modelling complex data distributions. By allowing each mixture component to capture a different part of the data distribution, mixture models are able to model non-linear relationships and account for multi-modal data. In addition, mixture models provide a natural way to incorporate unsupervised and semi-supervised learning, allowing for automatic feature and model selection.

Incorporating the AGGD and its bounded variant, BAGGD, as components of our proposed mixture models has proven to be useful in various energy management applications, including utility programs such as energy efficiency and demand response, as well as fault detection and diagnosis and occupancy estimation. The AGGD has a more flexible shape than the standard Gaussian distribution and can better model non-Gaussian and heavy-tailed data. Meanwhile, the bounded support of the BAGGD makes the incorporating mixture more robust to outliers and better suited for modelling energy consumption data that often have bounded support. By incorporating these distributions into our mixture models, we were able to accurately model complex energy consumption patterns. The proposed models have demonstrated their effectiveness in real-world scenarios and can help utilities and building managers make informed decisions regarding energy management and conservation efforts.

HMMs take this a step further by incorporating a temporal aspect into pattern recognition tasks. By modelling data as a sequence of observations, HMMs are able to capture the temporal dependencies between observations, making them particularly effective in speech recognition, handwriting recognition, and other sequential data analysis tasks. Mixture-based HMMs combine the strengths of both mixture models and HMMs, allowing for flexible modelling of complex data distributions with temporal dependencies. The number of hidden states in the HMM determines the complexity of the temporal dependencies, while the number of mixture components in each state captures the non-linear relationships within each state. This combination provides an effective approach to modelling complex data distributions in a variety of pattern recognition applications.

Considering our research into the potential of the bounded mixture models in the energy consumer characterization application, we concluded that our expectation-maximization algorithm using the MML criterion has successfully optimized the parameters of the bounded asymmetric generalized Gaussian mixture model and determined the optimal number of consumption profiles and the optimal subset of features simultaneously. This model has proven useful in categorizing households and how they consume energy in applications such as energy management, fault detection and diagnosis, and occupancy estimation. Our approach has outperformed all of the models used in the comparison in clustering synthetically generated smart meter records and real-life smart meter data. Additionally, it has accurately determined the optimal number of clusters in both real-life datasets. The proposed model has proven to be effective in determining households suitable for demand reduction initiatives, providing utility companies with environmentally friendly and cost-effective solutions. Future work can be done to develop a variational approach to learning the proposed model and a semi-supervised variant of our model.

Considering our research into the potential of the bounded mixture of mixtures in energy management applications, we proposed a novel statistical model that is a mixture of mixtures of bounded asymmetric generalized Gaussian and Uniform distributions and an unsupervised and semi-
-supervised learning approach using the EM algorithm to learn its parameters. We derived the objective function of the MML criterion in an attempt to discover the true number of components within the training dataset for the unsupervised learning approach. The proposed model outperforms all the state-of-the-art machine learning models used in the comparisons in both the semi-supervised and the unsupervised learning tasks in the three experiments with real-life data. The proposed component distribution is flexible and robust to outliers and can fit several asymmetric and bounded

support density distributions. Moreover, the model extension for explainability provides an accurate diagnosis of the detected patterns in fault detection applications in terms of the features used to train the model. In addition, the proposed workflow attempts to learn better models using low-level features and still provide explainability using high-level features. Overall, the proposed model is effective in determining households that are suitable for demand reduction initiatives, providing the opportunity for utility companies to adopt environmentally friendly and cost-effective technologies.

Considering our research in mixture-based hidden Markov models, we concluded that our proposed models, including the BAGGM-FSHMM, BAGGM-HMM, AGGM-FSHMM, and AGGM-HMM, demonstrated significant improvements in performance compared to other mixture-based HMMs in the field of HAR. Our experiments utilizing five sensor-based and two video-based HAR datasets showed that incorporating feature selection and bounded support distribution independently in a HAR system yields benefits, and incorporating these concepts simultaneously yields the best-performing model among the proposed models. Our top-performing model, the BAGGM-FSHMM, showed superior accuracy, precision, and F1-score results on all datasets, indicating that our proposed models led to an improvement that is demonstrated by all the performance measures we used in the experiments.

We focused our research on the HAR application to validate the performance of our proposed models, which we developed over years of research on challenging random variables. By demonstrating the effectiveness of our proposed models on a variety of datasets, we have shown that incorporating feature selection and bounded support distributions can improve the accuracy of machine learning models, particularly in datasets with high-dimensional features that have a dependence on the temporal axis. These findings can have significant implications for other real-world applications beyond HAR and provide a foundation for further research in this area.

In conclusion, this thesis has underscored the importance of developing a flexible and robust mixture model adept at managing asymmetric, non-Gaussian data distributions while mitigating the influence of outliers. The incorporation of bounded distributions, along with the Uniform distribution within an inner mixture, has proven to be a critical factor in crafting resilient models that effectively represent real-life data and resist distortion from noisy input. Furthermore, the significance of explainability in enhancing the credibility of deployed artificial intelligence models is paramount. Despite the proficiency of the models generated within this thesis, their interpretability by domain experts may be challenging. Thus, the integration of explainability is essential for providing straightforward, rule-based interpretations of the identified patterns. Mixture models, though valuable, exhibit increased efficacy when integrated with Hidden Markov Models (HMMs) to account for temporal dynamics. This combination is crucial for uncovering hidden patterns within the increasingly popular time series data, which is becoming a prerequisite for most real-life applications. Moreover, incorporating feature selection techniques within mixture models has proven to optimize accuracy while mitigating complexity. The inclusion of feature selection not only results in models that are robust against noisy and uninformative features but also yield computationally efficient models that can be deployed across a diverse range of hardware with varying capabilities. To propel this research forward, future endeavours should concentrate on amalgamating these multifaceted artificial intelligence concepts to devise versatile applications capable of efficiently learning from small sample populations, generalizing to wider contexts, minimizing computational complexity, maximizing accuracy in revealing hidden data structures and learning hyper-parameters automatically.

Appendix A

My Appendix

A.1

$$
\begin{aligned}
\frac{\partial \ln \Psi(X_{id}|\theta_{kd})}{\mu_{kd}} &= \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\mu_{kd}} & X_{id} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\mu_{kd}} & X_{id} \geq \mu_{kd} \end{cases} \\
&= \begin{cases} -A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kd}-X_{id})^{\lambda_{kd}-1}}{\sigma_{l_{kd}}^{\lambda_{kd}}} & X_{id} < \mu_{kd} \\ \\ A(\lambda_{kd})\lambda_{kd}\frac{(X_{id}-\mu_{kd})^{\lambda_{kd}-1}}{\sigma_{r_{kd}}^{\lambda_{kd}}} & X_{id} \geq \mu_{kd} \end{cases}
\end{aligned}
\tag{139}
$$

$$\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \mu_{kd}} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M) \times \begin{cases} \dfrac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}} + \\ \dfrac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}} du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} & x < \mu_{kd} \\ \\ \dfrac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}} + \\ \dfrac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}} du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} & x \geq \mu_{kd} \end{cases} \tag{140}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{kd})}{\partial \sigma_{l_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} & X_{id} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} & X_{id} \geq \mu_{kd} \end{cases}$$
$$= \begin{cases} A(\lambda_{kd}) \lambda_{kd} \frac{(\mu_{kd} - X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+1}} - \frac{1}{\sigma_{l_{kd}} + \sigma_{r_{kd}}} & X_{id} < \mu_{kd} \\ \\ -\frac{1}{\sigma_{l_{kd}} + \sigma_{r_{kd}}} & X_{id} \geq \mu_{kd} \end{cases} \tag{141}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{l_{kd}}} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M)$$
$$\times \begin{cases} \dfrac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} + \dfrac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} & x < \mu_{kd} \\ \\ \dfrac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} + \dfrac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma l_{kd}} du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} & x \geq \mu_{kd} \end{cases} \tag{142}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{kd})}{\partial \sigma_{r_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma_{r_{kd}}} & X_{id} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma_{r_{kd}}} & X_{id} \geq \mu_{kd} \end{cases}$$
$$= \begin{cases} -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} < \mu_{kd} \\ \\ \frac{A(\lambda_{kd})\lambda_{kd}}{\sigma_{r_{kd}}} \frac{(X_{id}-\mu_{kd})^{\lambda_{kd}}}{\sigma_{r_{kd}}^{\lambda_{kd}}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} \geq \mu_{kd} \end{cases} \tag{143}$$

$$\frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \sigma_{r_{kd}}} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M)$$
$$\times \begin{cases} \dfrac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma r_{kd}} + \dfrac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \sigma r_{kd}} du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} & x < \mu_{kd} \\ \\ \dfrac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma r_{kd}} + \dfrac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \sigma r_{kd}} du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} & x \geq \mu_{kd} \end{cases} \tag{144}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} & X_{id} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} & X_{id} \geq \mu_{kd} \end{cases}$$

$$= \begin{cases} \begin{aligned} &\frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ &\quad - \left(\frac{\mu_{kd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ &\quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) + \right. \\ &\quad \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ &\quad \left. + \ln\left(\frac{\mu_{kd} - X_{id}}{\sigma_{l_{kd}}}\right)\right] \end{aligned} & X_{id} < \mu_{kd} \\ \\ \begin{aligned} &\frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ &\quad - \left(\frac{X_{id} - \mu_{kd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ &\quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\ &\quad + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} + \\ &\quad \left. \ln\left(\frac{X_{id} - \mu_{kd}}{\sigma_{r_{kd}}}\right)\right] \end{aligned} & X_{id} \geq \mu_{kd} \end{cases} \tag{145}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \lambda_{kd}} &= \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M) \\ &\times \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} + \frac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} & x < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} + \frac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} & x \geq \mu_{kd} \end{cases} \end{aligned} \tag{146}$$

$$\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \mu_{kd}^2} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M)$$
$$\times \begin{cases} \begin{aligned} &\frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} + \frac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}} \right)^2 + \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} \right] du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} \\ &- \frac{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}} du \right)^2}{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd}) du \right)^2} \end{aligned} & x < \mu_{kd} \\ \\ \begin{aligned} &\frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} + \frac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}} \right)^2 + \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} \right] du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} \\ &- \frac{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}} du \right)^2}{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd}) du \right)^2} \end{aligned} & x \geq \mu_{kd} \end{cases} \tag{147}$$

$$\begin{aligned} \frac{\partial^2 \ln \Psi(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} &= \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} & X_{id} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \mu_{kd}^2} & X_{id} \geq \mu_{kd} \end{cases} \\ &= \begin{cases} -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1)\frac{(\partial \mu_{kd} - X_{id})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} < \mu_{kd} \\ \\ -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1)\frac{(X_{id} - \mu_{kd})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} \geq \mu_{kd} \end{cases} \end{aligned} \tag{148}$$

$$\frac{\partial^2\mathcal{L}(\mathcal{X},\Theta_M,Z,\varphi)}{\partial\sigma_{l_{kd}}^2}=\sum_{i=1}^{N}\frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}}p(k|\vec{X}_i,\Theta_M)$$

$$\times\begin{cases}\dfrac{\partial^2\ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2}+\dfrac{\int_{\partial k}g_1(X_{id}|\theta_{kd})\left[\left(\frac{\partial\ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}}\right)^2+\frac{\partial^2\ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2}\right]du}{\int_{\partial k}g_1(X_{id}|\theta_{kd})du} \\ -\dfrac{\left(\int_{\partial k}g_1(X_{id}|\theta_{kd})\frac{\partial\ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}}du\right)^2}{\left(\int_{\partial k}g_1(X_{id}|\theta_{kd})du\right)^2} & x<\mu_{kd} \\ \\ \dfrac{\partial^2\ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2}+\dfrac{\int_{\partial k}g_2(X_{id}|\theta_{kd})\left[\left(\frac{\partial\ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}}\right)^2+\frac{\partial^2\ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2}\right]du}{\int_{\partial k}g_2(X_{id}|\theta_{kd})du} \\ -\dfrac{\left(\int_{\partial k}g_2(X_{id}|\theta_{kd})\frac{\partial\ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}}du\right)^2}{\left(\int_{\partial k}g_2(X_{id}|\theta_{kd})du\right)^2} & x\geq\mu_{kd}\end{cases} \tag{149}$$

$$\frac{\partial^2\ln\Psi(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2}=\begin{cases}\frac{\partial^2\ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2} & X_{id}<\mu_{kd} \\ \\ \frac{\partial^2\ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{l_{kd}}^2} & X_{id}\geq\mu_{kd}\end{cases}$$
$$=\begin{cases}-(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kd}-X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}}+\frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id}<\mu_{kd} \\ \\ \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id}\geq\mu_{kd}\end{cases} \tag{150}$$

$$\frac{\partial^2 \mathcal{L}(\mathcal{X},\Theta_M,Z,\varphi)}{\partial\sigma^2_{r_{kd}}} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i,\Theta_M)$$

$$\times \begin{cases} \begin{aligned} &\frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}} + \frac{\int_{\partial k} g_1(X_{id}|\theta_{kd})\left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{r_{kd}}}\right)^2 + \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}}\right]du}{\int_{\partial k} g_1(X_{id}|\theta_{kd})du} \\ &- \frac{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd})\frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial\sigma_{r_{kd}}}du\right)^2}{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd})du\right)^2} \end{aligned} & x < \mu_{kd} \\ \\ \begin{aligned} &\frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}} + \frac{\int_{\partial k} g_2(X_{id}|\theta_{kd})\left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{r_{kd}}}\right)^2 + \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}}\right]du}{\int_{\partial k} g_2(X_{id}|\theta_{kd})du} \\ &- \frac{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd})\frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial\sigma_{r_{kd}}}du\right)^2}{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd})du\right)^2} \end{aligned} & x \geq \mu_{kd} \end{cases} \tag{151}$$

$$\begin{aligned} \frac{\partial^2 \ln \Psi(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}} &= \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}} & x < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial\sigma^2_{r_{kd}}} & x \geq \mu_{kd} \end{cases} \\ &= \begin{cases} \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & x < \mu_{kd} \\ \\ -(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(X_{id}-\mu_{kd})^{\lambda_{kd}}}{\sigma_{r_{kd}}^{\lambda_{kd}+2}} + \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & x \geq \mu_{kd} \end{cases} \end{aligned} \tag{152}$$

$$
\frac{\partial^2 \mathcal{L}(\mathcal{X}, \Theta_M, Z, \varphi)}{\partial \lambda_{kd}^2} = \sum_{i=1}^{N} \frac{\omega_d p(x_{id}|\theta_{kd})}{\zeta_{ikd}} p(k|\vec{X}_i, \Theta_M)
$$

$$
\times \begin{cases}
\dfrac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} + \dfrac{\int_{\partial k} g_1(X_{id}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} \right)^2 + \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\partial k} g_1(X_{id}|\theta_{kd}) du} \\
- \dfrac{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd}) \frac{\partial \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\partial k} g_1(X_{id}|\theta_{kd}) du \right)^2} & x < \mu_{kd} \\
\\
\dfrac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} + \dfrac{\int_{\partial k} g_2(X_{id}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} \right)^2 + \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\partial k} g_2(X_{id}|\theta_{kd}) du} \\
- \dfrac{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd}) \frac{\partial \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\partial k} g_2(X_{id}|\theta_{kd}) du \right)^2} & x \geq \mu_{kd}
\end{cases}
\tag{153}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} &= \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} & X_{id} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kd})}{\partial \lambda_{kd}^2} & X_{id} \geq \mu_{kd} \end{cases} \\
&= \begin{cases} -\dfrac{1}{\lambda_{kd}^2} + \dfrac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ -\dfrac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ -\left(\dfrac{\mu_{kd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \times \left[\dfrac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3}\right. & X_{id} < \mu_{kd} \\ + \left[\dfrac{1}{2} \ln\left(\dfrac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\ + \dfrac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ \left.\left. + \ln\left(\dfrac{\mu_{kd} - X_{id}}{\sigma_{l_{kd}}}\right)\right]^2\right] \\ \\ -\dfrac{1}{\lambda_{kd}^2} + \dfrac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ -\dfrac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ -\left(\dfrac{X_{id} - \mu_{kd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \times \left[\dfrac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3}\right. & X_{id} \geq \mu_{kd} \\ + \left[\dfrac{1}{2} \ln\left(\dfrac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\ + \dfrac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ \left.\left. + \ln\left(\dfrac{X_{id} - \mu_{kd}}{\sigma_{r_{kd}}}\right)\right]^2\right] \end{cases}
\end{aligned}
\tag{154}
$$

A.2

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \mu_{gd}} =
\begin{cases}
\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}} + \\
\frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} & X_{id} < \mu_{gd} \\
\\
\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}} + \\
\frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} & X_{id} \geq \mu_{gd}
\end{cases} \tag{155}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{gd})}{\mu_{gd}} = \begin{cases}
\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\mu_{gd}} & X_{id} < \mu_{gd} \\
\\
\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\mu_{gd}} & X_{id} \geq \mu_{gd}
\end{cases}$$
$$= \begin{cases}
-A(\lambda_{kd})\lambda_{kd} \frac{(\mu_{gd} - X_{id})^{\lambda_{kd}-1}}{\sigma_{l_{kd}}^{\lambda_{kd}}} & X_{id} < \mu_{gd} \\
\\
A(\lambda_{kd})\lambda_{kd} \frac{(X_{id} - \mu_{gd})^{\lambda_{kd}-1}}{\sigma_{r_{kd}}^{\lambda_{kd}}} & X_{id} \geq \mu_{gd}
\end{cases} \tag{156}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{l_{kd}}} =
\begin{cases}
\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} + \\
\frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} & X_{id} < \mu_{gd} \\
\\
\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} + \\
\frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} & X_{id} \geq \mu_{gd}
\end{cases} \tag{157}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma l_{kd}} & X_{id} < \mu_{gd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma l_{kd}} & X_{id} \geq \mu_{gd} \end{cases}$$

$$= \begin{cases} A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{gd}-X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+1}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} < \mu_{gd} \\ \\ -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} \geq \mu_{gd} \end{cases} \tag{158}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{r_{kd}}} =$$

$$\begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} + \\ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} & X_{id} < \mu_{gd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} + \\ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} & X_{id} \geq \mu_{gd} \end{cases} \tag{159}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} & X_{id} < \mu_{gd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} & X_{id} \geq \mu_{gd} \end{cases}$$

$$= \begin{cases} -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} < \mu_{gd} \\ \\ \frac{A(\lambda_{kd})\lambda_{kd}}{\sigma_{r_{kd}}} \frac{(X_{id}-\mu_{gd})^{\lambda_{kd}}}{\sigma_{r_{kd}}^{\lambda_{kd}}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} \geq \mu_{gd} \end{cases} \tag{160}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} + \\ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} & X_{id} < \mu_{gd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} + \\ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} & X_{id} \geq \mu_{gd} \end{cases} \tag{161}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} & X_{id} < \mu_{gd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} & X_{id} \geq \mu_{gd} \end{cases}$$
$$= \begin{cases} \frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ \quad - \left(\frac{\mu_{gd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) + \right. & X_{id} < \mu_{gd} \\ \quad \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ \quad \left. + \ln\left(\frac{\mu_{gd} - X_{id}}{\sigma_{l_{kd}}}\right)\right] \\ \\ \frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ \quad - \left(\frac{X_{id} - \mu_{gd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) \right. & X_{id} \geq \mu_{gd} \\ \quad + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} + \\ \quad \left. \ln\left(\frac{X_{id} - \mu_{gd}}{\sigma_{r_{kd}}}\right)\right] \end{cases} \tag{162}$$

$$
\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \mu_{gd}^2} =
\begin{cases}
\begin{aligned}
&\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \Big[\Big(\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}} \Big)^2 \Big] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&- \frac{\Big(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}} du \Big)^2}{\big(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du \big)^2}
\end{aligned} & X_{id} < \mu_{gd} \\
\\
\begin{aligned}
&\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \Big[\Big(\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}} \Big)^2 \Big] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&- \frac{\Big(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}} du \Big)^2}{\big(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du \big)^2}
\end{aligned} & X_{id} \geq \mu_{gd}
\end{cases}
\tag{163}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} &=
\begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} & X_{id} < \mu_{gd} \\
\\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \mu_{gd}^2} & X_{id} \geq \mu_{gd}
\end{cases} \\
&= \begin{cases}
-A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1)\frac{(\partial \mu_{gd} - X_{id})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} < \mu_{gd} \\
\\
-A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1)\frac{(X_{id} - \mu_{gd})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} \geq \mu_{gd}
\end{cases}
\end{aligned}
\tag{164}
$$

$$
\begin{aligned}
&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma^2_{l_{kd}}} \\
&= \begin{cases}
\begin{aligned}
&\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&- \frac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du \right)^2}
\end{aligned} & X_{id} < \mu_{gd} \\
\\
\begin{aligned}
&\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&- \frac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du \right)^2}
\end{aligned} & X_{id} \geq \mu_{gd}
\end{cases}
\end{aligned}
\tag{165}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} &= \begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} & X_{id} < \mu_{gd} \\
\\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} & X_{id} \geq \mu_{gd}
\end{cases} \\
&= \begin{cases}
-(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{gd}-X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} + \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id} < \mu_{gd} \\
\\
\frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id} \geq \mu_{gd}
\end{cases}
\end{aligned}
\tag{166}
$$

$$
\begin{aligned}
&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma^2_{r_{kd}}} \\
&= \begin{cases}
\dfrac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{r_{kd}}} \\
+ \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
+ \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} & X_{id} < \mu_{gd} \\
- \dfrac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du \right)^2} \\
\\
\dfrac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{r_{kd}}} \\
+ \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
+ \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} & X_{id} \geq \mu_{gd} \\
- \dfrac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du \right)^2}
\end{cases}
\end{aligned}
\tag{167}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} &= \begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} & X_{id} < \mu_{gd} \\
\\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \sigma^2_{l_{kd}}} & X_{id} \geq \mu_{gd}
\end{cases} \\
&= \begin{cases}
-(\lambda_{kd}+1) A(\lambda_{kd}) \lambda_{kd} \frac{(\mu_{gd} - X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} \\
+ \frac{1}{(\sigma_{l_{kd}} + \sigma_{r_{kd}})^2} & X_{id} < \mu_{gd} \\
\\
\frac{1}{(\sigma_{l_{kd}} + \sigma_{r_{kd}})^2} & X_{id} \geq \mu_{gd}
\end{cases}
\end{aligned}
\tag{168}
$$

$$
\begin{aligned}
&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}^2} \\
&= \begin{cases}
\begin{aligned}
&\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du} \\
&- \frac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) \frac{\partial \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{gd}) du \right)^2}
\end{aligned} & X_{id} < \mu_{gd} \\
\\
\begin{aligned}
&\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du} \\
&- \frac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) \frac{\partial \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{gd}) du \right)^2}
\end{aligned} & X_{id} \geq \mu_{gd}
\end{cases}
\end{aligned}
\tag{169}
$$

$$
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} = \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} & X_{id} < \mu_{gd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{gd})}{\partial \lambda_{kd}^2} & X_{id} \geq \mu_{gd} \end{cases}
$$

$$
= \begin{cases} \begin{array}{l} -\dfrac{1}{\lambda_{kd}^2} + \dfrac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ -\dfrac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ -\left(\dfrac{\mu_{gd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \times \Bigg[\dfrac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3} \\ + \Bigg[\dfrac{1}{2} \ln\left(\dfrac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) \\ + \dfrac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ + \ln\left(\dfrac{\mu_{gd} - X_{id}}{\sigma_{l_{kd}}}\right) \Bigg]^2 \Bigg] \end{array} & X_{id} < \mu_{gd} \\ \\ \begin{array}{l} -\dfrac{1}{\lambda_{kd}^2} + \dfrac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ -\dfrac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ -\left(\dfrac{X_{id} - \mu_{gd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \times \Bigg[\dfrac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3} \\ + \Bigg[\dfrac{1}{2} \ln\left(\dfrac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) \\ + \dfrac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ + \ln\left(\dfrac{X_{id} - \mu_{gd}}{\sigma_{r_{kd}}}\right) \Bigg]^2 \Bigg] \end{array} & X_{id} \geq \mu_{gd} \end{cases} \tag{170}
$$

A.3

$$\frac{\partial \log \phi(\vec{X}_i|\theta_{kmd})}{\partial \mu_{kmd}} = \begin{cases} \dfrac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} + \\ \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\ \\ \dfrac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} + \\ \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \end{cases} \tag{171}$$

$$\begin{aligned} \frac{\partial \ln \Psi(X_{id}|\theta_{kmd})}{\mu_{kmd}} &= \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\mu_{kmd}} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\mu_{kmd}} & X_{id} \geq \mu_{kmd} \end{cases} \\ &= \begin{cases} -A(\lambda_{kd})\lambda_{kd} \frac{(\mu_{kmd} - X_{id})^{\lambda_{kd}-1}}{\sigma_{l_{kd}}^{\lambda_{kd}}} & X_{id} < \mu_{kmd} \\ \\ A(\lambda_{kd})\lambda_{kd} \frac{(X_{id} - \mu_{kmd})^{\lambda_{kd}-1}}{\sigma_{r_{kd}}^{\lambda_{kd}}} & X_{id} \geq \mu_{kmd} \end{cases} \end{aligned} \tag{172}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{l_{kd}}} = \begin{cases} \dfrac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} + \\ \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\ \\ \dfrac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} + \\ \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \end{cases} \tag{173}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma l_{kd}} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma l_{kd}} & X_{id} \geq \mu_{kmd} \end{cases}$$
$$= \begin{cases} A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kmd}-X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+1}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} < \mu_{kmd} \\ \\ -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} \geq \mu_{kmd} \end{cases} \tag{174}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{r_{kd}}} =$$
$$\begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} + \\ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd})\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} + \\ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd})\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \end{cases} \tag{175}$$

$$\frac{\partial \ln \Psi(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} & X_{id} \geq \mu_{kmd} \end{cases}$$
$$= \begin{cases} -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} < \mu_{kmd} \\ \\ \frac{A(\lambda_{kd})\lambda_{kd}}{\sigma_{r_{kd}}}\frac{(X_{id}-\mu_{kmd})^{\lambda_{kd}}}{\sigma_{r_{kd}}^{\lambda_{kd}}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{id} \geq \mu_{kmd} \end{cases} \tag{176}$$

$$
\begin{aligned}
&\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}} = \\
&\begin{cases}
\dfrac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} + \\
\dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\
\\
\dfrac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} + \\
\dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd}
\end{cases}
\end{aligned}
\tag{177}
$$

$$
\begin{aligned}
\frac{\partial \ln \Psi(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} &= \begin{cases}
\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} & X_{id} < \mu_{kmd} \\
\\
\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} & X_{id} \geq \mu_{kmd}
\end{cases} \\
&= \begin{cases}
\begin{aligned}
&\frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\
&\quad - \left(\frac{\mu_{kmd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\
&\quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) + \right. \\
&\quad \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\
&\quad \left. + \ln\left(\frac{\mu_{kmd} - X_{id}}{\sigma_{l_{kd}}}\right)\right]
\end{aligned} & X_{id} < \mu_{kmd} \\
\\
\begin{aligned}
&\frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\
&\quad - \left(\frac{X_{id} - \mu_{kmd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\
&\quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\
&\quad + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} + \\
&\quad \left. \ln\left(\frac{X_{id} - \mu_{kmd}}{\sigma_{r_{kd}}}\right)\right]
\end{aligned} & X_{id} \geq \mu_{kmd}
\end{cases}
\end{aligned}
\tag{178}
$$

$$\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \mu_{kmd}^2} = \begin{cases} \dfrac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} \\ + \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} \\ + \dfrac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\ - \dfrac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du \right)^2} \\ \\ \dfrac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} \\ + \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} \\ + \dfrac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \\ - \dfrac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du \right)^2} \end{cases} \tag{179}$$

$$\begin{aligned} \frac{\partial^2 \ln \Psi(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} &= \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \mu_{kmd}^2} & X_{id} \geq \mu_{kmd} \end{cases} \\ &= \begin{cases} -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1) \frac{(\partial \mu_{kmd} - X_{id})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} < \mu_{kmd} \\ \\ -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd} - 1) \frac{(X_{id} - \mu_{kmd})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{id} \geq \mu_{kmd} \end{cases} \end{aligned} \tag{180}$$

$$
\begin{aligned}
&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g,\sigma_{l_g},\sigma_{r_g},\lambda_g)}{\partial \sigma^2_{l_{kd}}} \\
&= \begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & \\
+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & \\
+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\
- \frac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du \right)^2} & \\
 & \\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & \\
+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & \\
+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \\
- \frac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du \right)^2} &
\end{cases}
\end{aligned}
\tag{181}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} &= \begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & X_{id} < \mu_{kmd} \\
 & \\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & X_{id} \geq \mu_{kmd}
\end{cases} \\
&= \begin{cases}
-(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kmd}-X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} + \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id} < \mu_{kmd} \\
 & \\
\frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{id} \geq \mu_{kmd}
\end{cases}
\end{aligned}
\tag{182}
$$

$$
\begin{aligned}
&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma^2_{r_{kd}}} \\
&= \begin{cases}
\begin{aligned}
&\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{r_{kd}}} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} \\
&+ \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} \\
&- \frac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du \right)^2}
\end{aligned} & X_{id} < \mu_{kmd} \\
\\
\begin{aligned}
&\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{r_{kd}}} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} \\
&+ \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} \\
&- \frac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du \right)^2}
\end{aligned} & X_{id} \geq \mu_{kmd}
\end{cases}
\end{aligned}
\tag{183}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \Psi(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} &= \begin{cases}
\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & X_{id} < \mu_{kmd} \\
\\
\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \sigma^2_{l_{kd}}} & X_{id} \geq \mu_{kmd}
\end{cases} \\
&= \begin{cases}
\begin{aligned}
&-(\lambda_{kd}+1) A(\lambda_{kd}) \lambda_{kd} \tfrac{(\mu_{kmd} - X_{id})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} \\
&+ \tfrac{1}{(\sigma_{l_{kd}} + \sigma_{r_{kd}})^2}
\end{aligned} & X_{id} < \mu_{kmd} \\
\\
\tfrac{1}{(\sigma_{l_{kd}} + \sigma_{r_{kd}})^2} & X_{id} \geq \mu_{kmd}
\end{cases}
\end{aligned}
\tag{184}
$$

$$\begin{aligned}&\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}^2}\\ &= \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} \\ \quad + \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} \\ \quad + \frac{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du} & X_{id} < \mu_{kmd} \\ \quad - \frac{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) \frac{\partial \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{id}|\theta_{kmd}) du \right)^2} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} \\ \quad + \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\left(\frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} \\ \quad + \frac{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \left[\frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du} & X_{id} \geq \mu_{kmd} \\ \quad - \frac{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) \frac{\partial \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{id}|\theta_{kmd}) du \right)^2} \end{cases}\end{aligned} \tag{185}$$

$$\frac{\partial^2 \ln \Psi(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} = \begin{cases} \frac{\partial^2 \ln g_1(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} & X_{id} < \mu_{kmd} \\ \\ \frac{\partial^2 \ln g_2(X_{id}|\theta_{kmd})}{\partial \lambda_{kd}^2} & X_{id} \geq \mu_{kmd} \end{cases}$$

$$= \begin{cases} \begin{aligned} & - \frac{1}{\lambda_{kd}^2} + \frac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ & - \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ & - \left(\frac{\mu_{kmd} - X_{id}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ & \times \Bigg[\frac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3} \\ & + \Bigg[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) \\ & + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ & + \ln\left(\frac{\mu_{kmd} - X_{id}}{\sigma_{l_{kd}}}\right) \Bigg]^2 \Bigg] \end{aligned} & X_{id} < \mu_{kmd} \\ \\ \begin{aligned} & - \frac{1}{\lambda_{kd}^2} + \frac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ & - \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ & - \left(\frac{X_{id} - \mu_{kmd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ & \times \Bigg[\frac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3} \\ & + \Bigg[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) \\ & + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ & + \ln\left(\frac{X_{id} - \mu_{kmd}}{\sigma_{r_{kd}}}\right) \Bigg]^2 \Bigg] \end{aligned} & X_{id} \geq \mu_{kmd} \end{cases} \tag{186}$$

A.4

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \mu_{kd}} = \begin{cases} \begin{array}{l} \dfrac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} + \\ \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \end{array} & X_{dt} < \mu_{kd} \\ \\ \begin{array}{l} \dfrac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} + \\ \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \end{array} & X_{dt} \geq \mu_{kd} \end{cases} \tag{187}$$

$$\begin{aligned} \frac{\partial \ln \Psi(X_{dt}|\theta_{kd})}{\mu_{kd}} &= \begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\mu_{kd}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\mu_{kd}} & X_{dt} \geq \mu_{kd} \end{cases} \\ &= \begin{cases} -A(\lambda_{kd})\lambda_{kd} \frac{(\mu_{kd} - X_{dt})^{\lambda_{kd}-1}}{\sigma_{l_{kd}}^{\lambda_{kd}}} & X_{dt} < \mu_{kd} \\ \\ A(\lambda_{kd})\lambda_{kd} \frac{(X_{dt} - \mu_{kd})^{\lambda_{kd}-1}}{\sigma_{r_{kd}}^{\lambda_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases} \end{aligned} \tag{188}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{l_{kd}}} = \begin{cases} \begin{array}{l} \dfrac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} + \\ \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \end{array} & X_{dt} < \mu_{kd} \\ \\ \begin{array}{l} \dfrac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} + \\ \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \end{array} & X_{dt} \geq \mu_{kd} \end{cases} \tag{189}$$

$$\frac{\partial \ln \Psi(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma l_{kd}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma l_{kd}} & X_{dt} \geq \mu_{kd} \end{cases}$$

$$= \begin{cases} A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kd}-X_{dt})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+1}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{dt} < \mu_{kd} \\ \\ -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases} \tag{190}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma_{r_{kd}}} =$$

$$\begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} + \\ \frac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd})\frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} + \\ \frac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd})\frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \end{cases} \tag{191}$$

$$\frac{\partial \ln \Psi(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} = \begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases}$$

$$= \begin{cases} -\frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{dt} < \mu_{kd} \\ \\ \frac{A(\lambda_{kd})\lambda_{kd}}{\sigma_{r_{kd}}}\frac{(X_{dt}-\mu_{kd})^{\lambda_{kd}}}{\sigma_{r_{kd}}^{\lambda_{kd}}} - \frac{1}{\sigma_{l_{kd}}+\sigma_{r_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases} \tag{192}$$

$$\frac{\partial \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}} = \begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} + \\ \frac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} + \\ \frac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \end{cases} \tag{193}$$

$$\frac{\partial \ln \Psi(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} = \begin{cases} \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} & X_{dt} \geq \mu_{kd} \end{cases}$$

$$= \begin{cases} \frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ \quad - \left(\frac{\mu_{kd} - X_{dt}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right) + \right. & X_{dt} < \mu_{kd} \\ \quad \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ \quad \left. + \ln\left(\frac{\mu_{kd} - X_{dt}}{\sigma_{l_{kd}}}\right)\right] \\ \\ \frac{1}{\lambda_{kd}} + \frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{2\lambda_{kd}^2} \\ \quad - \left(\frac{X_{dt} - \mu_{kd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ \quad \times \left[\frac{1}{2} \ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. & X_{dt} \geq \mu_{kd} \\ \quad + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} + \\ \quad \left. \ln\left(\frac{X_{dt} - \mu_{kd}}{\sigma_{r_{kd}}}\right)\right] \end{cases} \tag{194}$$

$$\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \mu_{kd}^2} =$$

$$\begin{cases} \dfrac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du \right)^2} \\ \\ \dfrac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du \right)^2} \end{cases} \tag{195}$$

$$\begin{aligned} \frac{\partial^2 \ln \Psi(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} &= \begin{cases} \frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \mu_{kd}^2} & X_{dt} \geq \mu_{kd} \end{cases} \\ &= \begin{cases} -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd}-1)\frac{(\partial \mu_{kd} - X_{dt})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{dt} < \mu_{kd} \\ \\ -A(\lambda_{kd})\lambda_{kd}(\lambda_{kd}-1)\frac{(X_{dt} - \mu_{kd})^{\lambda_{kd}-2}}{\sigma_{kd}^{\lambda_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases} \end{aligned} \tag{196}$$

$$\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma^2_{l_{kd}}}$$

$$= \begin{cases} \dfrac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du \right)^2} \\ \\ \dfrac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{l_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du \right)^2} \end{cases} \tag{197}$$

$$\frac{\partial^2 \ln \Psi(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} = \begin{cases} \frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases}$$
$$= \begin{cases} -(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kd}-X_{dt})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} + \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{dt} < \mu_{kd} \\ \\ \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{dt} \geq \mu_{kd} \end{cases} \tag{198}$$

$$\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \sigma^2_{r_{kd}}}$$

$$= \begin{cases} \dfrac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{r_{kd}}} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du \right)^2} \\ \\ \dfrac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{r_{kd}}} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{r_{kd}}} \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma_{r_{kd}}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du \right)^2} \end{cases} \tag{199}$$

$$\frac{\partial^2 \ln \Psi(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} = \begin{cases} \frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \sigma^2_{l_{kd}}} & X_{dt} \geq \mu_{kd} \end{cases}$$

$$= \begin{cases} -(\lambda_{kd}+1)A(\lambda_{kd})\lambda_{kd}\frac{(\mu_{kd}-X_{dt})^{\lambda_{kd}}}{\sigma_{l_{kd}}^{\lambda_{kd}+2}} \\ +\frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{dt} < \mu_{kd} \\ \\ \frac{1}{(\sigma_{l_{kd}}+\sigma_{r_{kd}})^2} & X_{dt} \geq \mu_{kd} \end{cases} \tag{200}$$

$$\frac{\partial^2 \log \phi(\vec{X}_i|\mu_g, \sigma_{l_g}, \sigma_{r_g}, \lambda_g)}{\partial \lambda_{kd}^2}$$

$$= \begin{cases} \dfrac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du} & X_{dt} < \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) \frac{\partial \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_1(X_{dt}|\theta_{kd}) du \right)^2} \\ \\ \dfrac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\left(\frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} \right)^2 \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} \\ + \dfrac{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \left[\frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} \right] du}{\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du} & X_{dt} \geq \mu_{kd} \\ - \dfrac{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) \frac{\partial \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}} du \right)^2}{\left(\int_{\tau_g} g_2(X_{dt}|\theta_{kd}) du \right)^2} \end{cases} \tag{201}$$

$$
\frac{\partial^2 \ln \Psi(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} = \begin{cases} \frac{\partial^2 \ln g_1(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} & X_{dt} < \mu_{kd} \\ \\ \frac{\partial^2 \ln g_2(X_{dt}|\theta_{kd})}{\partial \lambda_{kd}^2} & X_{dt} \geq \mu_{kd} \end{cases}
$$

$$
= \begin{cases} \begin{aligned} & -\frac{1}{\lambda_{kd}^2} + \frac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ & -\frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ & -\left(\frac{\mu_{kd} - X_{dt}}{\sigma_{l_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ & \times \left[\frac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3}\right. \\ & + \left[\frac{1}{2}\ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\ & + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ & \left.\left. + \ln\left(\frac{\mu_{kd} - X_{dt}}{\sigma_{l_{kd}}}\right)\right]^2\right] \end{aligned} & X_{dt} < \mu_{kd} \\ \\ \begin{aligned} & -\frac{1}{\lambda_{kd}^2} + \frac{3(\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^4} \\ & -\frac{3(\psi(1/\lambda_{kd}) - \psi(3/\lambda_{kd}))}{\lambda_{kd}^3} \\ & -\left(\frac{X_{dt} - \mu_{kd}}{\sigma_{r_{kd}}}\right)^{\lambda_{kd}} A(\lambda_{kd}) \\ & \times \left[\frac{(9\psi'(3/\lambda_{kd}) - \psi'(1/\lambda_{kd}))}{2\lambda_{kd}^3}\right. \\ & + \left[\frac{1}{2}\ln\left(\frac{\Gamma(3/\lambda_{kd})}{\Gamma(1/\lambda_{kd})}\right)\right. \\ & + \frac{(\psi(1/\lambda_{kd}) - 3\psi(3/\lambda_{kd}))}{2\lambda_{kd}} \\ & \left.\left. + \ln\left(\frac{X_{dt} - \mu_{kd}}{\sigma_{r_{kd}}}\right)\right]^2\right] \end{aligned} & X_{dt} \geq \mu_{kd} \end{cases} \tag{202}
$$

Bibliography

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, and et al., “A public domain dataset for human activity recognition using smartphones,” in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, vol. 3, (Bruges, Belgium), pp. 3–10, 2013.

[2] S. Kung, M. Mak, and S. Lin, *Biometric Authentication: A Machine Learning Approach*. Prentice Hall Information and System Sciences Series, Prentice Hall, 1 ed., 2004.

[3] E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: An application to face detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Juan, Puerto Rico, USA), pp. 130–136, 1997.

[4] A. Jain, R. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[5] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 84–95, 1980.

[6] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies, “Detection of forgery in paintings using supervised learning,” in *IEEE International Conference on Image Processing (ICIP)*, pp. 2921–2924, 2009.

[7] M. Dorigo and U. Schnepf, “Genetics-based machine learning and behaviour based robotics: A new synthesis,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 1, pp. 141–154, 1993.

[8] B. Delcroix, J. L. Ny, M. Bernier, M. Azam, B. Qu, and J.-S. Venne, “Autoregressive neural networks with exogenous variables for indoor temperature prediction in buildings,” in *Building Simulation*, vol. 14, pp. 165–178, Springer, 2021.

[9] O. Graja, M. Azam, and N. Bouguila, “Breast cancer diagnosis using quality control charts and logistic regression,” in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 215–220, IEEE, 2018.

[10] M. Azam and N. Bouguila, “Speaker classification via supervised hierarchical clustering using ica mixture model,” in *Image and Signal Processing: 7th International Conference, ICISP 2016, Trois-Rivières, QC, Canada, May 30-June 1, 2016, Proceedings 7*, pp. 193–202, Springer, 2016.

[11] I. Channoufi, F. Najar, S. Bourouis, M. Azam, A. S. Halibas, R. Alroobaea, and A. Al-Badi, “Flexible statistical learning model for unsupervised image modeling and segmentation,” *Mixture Models and Applications*, pp. 325–348, 2020.

[12] M. Azam, J. P. Singh, and N. Bouguila, “Spatial image segmentation based on beta-liouville mixture models and markov random field,” in *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, vol. 1, pp. 936–941, IEEE, 2021.

[13] A. Algumaei, M. Azam, F. Najar, and N. Bouguila, “Bounded multivariate generalized gaussian mixture model using ica and iva,” *Pattern Analysis and Applications*, pp. 1–30, 2023.

[14] M. Azam and N. Bouguila, “Blind source separation as pre-processing to unsupervised keyword spotting via an ica mixture model,” in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 833–836, IEEE, 2018.

[15] T. Elguebaly and N. Bouguila, “Simultaneous high-dimensional clustering and feature selection using asymmetric gaussian mixture models,” *Image and Vision Computing*, vol. 34, pp. 27–41, 2015.

[16] M. Allili, N. Bouguila, and D. A. Ziou, “Robust video foreground segmentation by using generalized gaussian mixture modeling,” in *Fourth Canadian Conference on Computer and Robot Vision (CRV’07)*, pp. 503–509, IEEE, 2007.

[17] C. Rasmussen, “The infinite gaussian mixture model,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 554–560, 1999.

[18] C. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007.

[19] D. Ap, “Maximum likelihood from incomplete data via em algorithm,” *J. Royal Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[20] M. Yang, C. Lai, and C. A. Lin, “robust em clustering algorithm for Gaussian mixture models,” *Pattern Recognition*, vol. 45, pp. 3950–3961, 2012.

[21] T. M. Nguyen, Q. J. Wu, and H. Zhang, “Bounded generalized gaussian mixture model,” *Pattern Recognition*, vol. 47, no. 9, pp. 3132–3142, 2014.

[22] M. Azam and N. Bouguila, “Multivariate-bounded gaussian mixture model with minimum message length criterion for model selection (2021),” *Expert Systems*, vol. 38, no. 5, p. e12688, 2021.

[23] M. Azam and N. Bouguila, “Unsupervised keyword spotting using bounded generalized gaussian mixture model with ica,” in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1150–1154, IEEE, 2015.

[24] A. A. Farag, A. S. El-Baz, and G. Gimel’farb, “Precise segmentation of multimodal images,” *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 952–968, 2006.

[25] P. Hedelin and J. Skoglund, “Vector quantization based on gaussian mixture models,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 385–401, 2000.

[26] J. Lindblom and J. Samuelsson, “Bounded support gaussian mixture modeling of speech spectra,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 1, pp. 88–99, 2003.

[27] J. S. Kalsi, M. Azam, and N. Bouguila, "Color image segmentation using semi-bounded finite mixture models by incorporating mean templates," *Mixture Models and Applications*, pp. 273–305, 2020.

[28] K. Kokkinakis and A. Nandi, "Exponent parameter estimation for generalized Gaussian probability density functions with application to speech modeling," *Signal Processing*, vol. 85, pp. 1852–1858, 2005.

[29] J. Banfield and A. Raftery, "Model-based Gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, 1993.

[30] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2041–2050, ACM, 2018.

[31] M. Sadooghi and S. Khadem, "Improving one class support vector machine novelty detection scheme using nonlinear features," *Pattern Recognition*, vol. 83, pp. 14–33, 2018.

[32] T. Kieu, B. Yang, C. Guo, and C. Jensen, "Outlier detection for time series with recurrent autoencoder ensembles," *IJCAI*, pp. 2725–2732, 2019.

[33] M. Goldstein and A. Dengel, "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," in *KI-2012: Poster and Demo Track*, vol. 9, 2012.

[34] M. Azam and N. Bouguila, "Texture image categorization in wavelet domain via naive bayes classifier based on laplace and generalized gaussian distribution," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 143–150, IEEE, 2019.

[35] M. Palacios and M. Steel, "Non-gaussian bayesian geostatistical modeling," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 604–618, 2006.

[36] J. Miller and J. Thomas, "Detectors for discrete-time signals in non-gaussian noise," *IEEE Transactions on Information Theory*, vol. 18, no. 2, pp. 241–250, 1972.

[37] N. Farvardin and J. Modestino, "Optimum quantizer performance for a class of non-gaussian memoryless sources," *IEEE Transactions on Information Theory*, vol. 30, no. 3, pp. 485–497, 1984.

[38] T. Elguebaly and N. Bouguila, "Bayesian learning of finite generalized gaussian mixture models on images," *Signal Processing*, vol. 91, no. 4, pp. 801–820, 2011.

[39] W. Mauersberger, "Experimental results on the performance of mismatched quantizers," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 381–386, 1979.

[40] M. S. Allili, N. Bouguila, and D. Ziou, "Finite general gaussian mixture modeling and application to image and video foreground segmentation," *Journal of Electronic Imaging*, vol. 17, no. 1, p. 013005, 2008.

[41] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[42] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[43] M. Azam, *Bounded Support Finite Mixtures for Multidimensional Data Modeling and Clustering*. PhD thesis, Concordia University, 2019.

[44] M. Azam and N. Bouguila, "Multivariate bounded support laplace mixture model," *Soft Computing*, vol. 24, no. 17, pp. 13239–13268, 2020.

[45] O. Dalhoumi, M. Amayri, and N. Bouguila, "A review of neural networks for buildings occupancy measurement," in *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 29–35, 2022.

[46] Z. Xian, M. Azam, M. Amayri, and N. Bouguila, "Model selection criterion for multivariate bounded asymmetric gaussian mixture model," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1436–1440, IEEE, 2021.

[47] H. Nguyen, K. Maanicshah, M. Azam, and N. Bouguila, “Data clustering using variational learning of finite scaled dirichlet mixture models with component splitting,” in *Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part II 16*, pp. 117–128, Springer, 2019.

[48] H. Nguyen, M. Azam, and N. Bouguila, “Data clustering using variational learning of finite scaled dirichlet mixture models,” in *2019 IEEE 28th international symposium on industrial electronics (ISIE)*, pp. 1391–1396, IEEE, 2019.

[49] K. Maanicshah, M. Azam, H. Nguyen, N. Bouguila, and W. Fan, “Finite inverted beta-liouville mixture models with variational component splitting,” *Mixture Models and Applications*, pp. 209–233, 2020.

[50] H. Nguyen, M. Kalra, M. Azam, and N. Bouguila, “Data clustering using online variational learning of finite scaled dirichlet mixture models,” in *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)*, pp. 267–274, IEEE, 2019.

[51] R. T. Vemuri, M. Azam, Z. Patterson, and N. Bouguila, “Bayesian inference of hidden markov models using dirichlet mixtures,” in *Hidden Markov Models and Applications*, pp. 157–176, Springer, 2012.

[52] R. T. Vemuri, M. Azam, N. Bouguila, and Z. Patterson, “A bayesian sampling framework for asymmetric generalized gaussian mixture models learning,” *Neural Computing and Applications*, pp. 1–12, 2021.

[53] S. Bourouis, N. Bouguila, Y. Li, and M. Azam, “Visual scene reconstruction using a bayesian learning framework,” in *Image and Signal Processing: 8th International Conference, ICISP 2018, Cherbourg, France, July 2-4, 2018, Proceedings 8*, pp. 225–232, Springer, 2018.

[54] R. T. Vemuri, M. Azam, N. Bouguila, and Z. Patterson, “Bayesian model and feature selection in asymmetric generalized gaussian mixtures,” in *2022 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1–6, IEEE, 2022.

[55] R. Khardon and D. Roth, "Model-based subspace clustering," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 169–193, 1997.

[56] B. Alghabashi, M. Al Mashrgy, M. Azam, and N. Bouguila, "Finite multi-dimensional generalized gamma mixture model learning for feature selection," in *Learning Control*, pp. 147–173, Elsevier, 2021.

[57] C. Cardie, "A cognitive bias approach to feature selection and weighting for case-based learners," *Machine Learning*, vol. 41, no. 1, pp. 85–116, 2000.

[58] B. Heisele, T. Serre, S. Mukherjee, and T. Poggio, "Feature reduction and hierarchy of classifiers for fast object detection in video images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18–24, 2001.

[59] K. Kümmel, T. Scheidat, C. Vielhauer, and J. Dittmann, "Feature selection on handwriting biometrics: security aspects of artificial forgeries," in *International Conference on Communications and Multimedia Security (CMS)*, pp. 16–25, 2012.

[60] S. Boutemedjet, N. Bouguila, and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, 2009.

[61] H.-L. Wei and S. Billings, "Feature subset selection and ranking for data dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 162–166, 2007.

[62] M. H. Law, M. A. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[63] T. Elguebaly and N. Bouguila, "Model-based approach for high-dimensional non-gaussian visual data clustering and feature weighting," *Digital Signal Processing*, vol. 40, pp. 63–79, 2015.

[64] Z. Song, S. Ali, and N. Bouguila, “Background subtraction using infinite asymmetric gaussian mixture models with simultaneous feature selection,” *IET Image Processing*, vol. 14, no. 11, pp. 2321–2332, 2020.

[65] Z. Song, S. Ali, and N. Bouguila, “Bayesian inference for infinite asymmetric gaussian mixture with feature selection,” *Soft Computing*, vol. 25, no. 12, pp. 6043–6053, 2021.

[66] W. Pan and X. Shen, “Penalized model-based clustering with application to variable selection,” *Journal of Machine Learning Research*, vol. 8, pp. 1145–1164, 2007.

[67] Z. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, pp. 31–57, 2018.

[68] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable machine learning–a brief history, state-of-the-art and challenges,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 417–431, Springer, 2020.

[69] W. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. I. m. l. d. Yu, “methods, and applications,” tech. rep., *ArXiv*, type = Preprint, archivePrefix = arXiv, eprint = 1901.04592, 2019.

[70] S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian, “Explainable k-means clustering: Theory and practice,” in *XXAI Workshop. ICML*, 2020.

[71] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[72] K. Prabhakaran, J. Dridi, M. Amayri, and N. Bouguila, “Explainable k-means clustering for occupancy estimation,” *Procedia Computer Science*, vol. 203, pp. 326–333, 2022.

[73] S. Haben, C. Singleton, and P. Grindrod, “Analysis and clustering of residential customers energy behavioral demand using smart meter data,” *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136–144, 2015.

[74] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[75] P. Faria, J. Spinola, and Z. Vale, “Aggregation and remuneration of electricity consumers and producers for the definition of demand-response programs,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 952–961, 2016.

[76] D. Li, W.-Y. Chiu, H. Sun, and H. V. Poor, “Multiobjective optimization for demand side management program in smart grid,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1482–1490, 2017.

[77] N. Al Khafaf, M. Jalili, and P. Sokolowski, “Application of deep learning long short-term memory in energy demand forecasting,” in *International Conference on Engineering Applications of Neural Networks*, pp. 31–42, Springer, 2019.

[78] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[79] L. Rabiner and B. Juang, “An introduction to hidden markov models,” *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[80] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose, “Fine-grained activity recognition by aggregating abstract object usage,” in *Ninth IEEE International Symposium on Wearable Computers (ISWC’05)*, pp. 44–51, IEEE, 2005.

[81] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[82] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, “One-shot learning of human activity with an map adapted gmm and simplex-hmm,” *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1769–1780, 2017.

[83] W. Wang, K.-T. Chen, and W.-C. Li, “Deep learning for sensor-based activity recognition: A survey,” in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 1067–1074, IEEE, 2018.

[84] N. Manouchehri and N. Bouguila, “Integration of multivariate beta-based hidden markov models and support vector machines with medical applications,” in *Proceedings of the International FLAIRS Conference Proceedings*, vol. 35, (Hutchinson Island, FL, USA), pp. 15–18, 2022.

[85] R. Nasfi and N. Bouguila, “Indoor activity recognition using a hybrid generative-discriminative approach with hidden markov models and support vector machines,” in *2022 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1–6, 2022.

[86] J. A. Hartigan, *Clustering Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 1975.

[87] S. Ali and N. Bouguila, “A roadmap to hidden markov models and a review of its application in occupancy estimation,” in *Hidden Markov Models and Applications*, pp. 1–31, InTech, 2012.

[88] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[89] CER, “CER smart metering project - electricity customer behaviour trial, 2009-2010 [dataset].” `https://www.esb.ie/docs/default-source/docs-nosales/cer-smart-metering-project-electricity-customer-behaviour-\trial-2009-2010.pdf`, 2012. Accessed: Apr. 5, 2023.

[90] S. Haben, J. Ward, D. V. Greetham, C. Singleton, and P. Grindrod, “A new error measure for forecasts of household-level, high resolution electrical energy consumption,” *International Journal of Forecasting*, vol. 30, no. 2, pp. 246–256, 2014.

[91] H.-A. Cao, C. Beckel, and T. Staake, “Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns,” in *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*, pp. 4733–4738, IEEE, 2013.

[92] U. P. Networks, “Smartmeter energy consumption data in london households, 2011-2014 [dataset],” 2013.

[93] U. of Massachusetts Amherst, “Umass smart* dataset - microgrid dataset, 2013 release [dataset],” 2013.

[94] D. Alahakoon and X. Yu, “Smart electricity meter data intelligence for future energy systems: A survey,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 425–436, 2015.

[95] N. Al Khafaf, M. Jalili, and P. Sokolowski, “Demand response planning tool using markov decision process,” in *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pp. 484–489, IEEE, 2018.

[96] A. Shahzadeh, A. Khosravi, and S. Nahavandi, “Improving load forecast accuracy by clustering consumers using smart meter data,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2015.

[97] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer characterization options for improving the tariff offer,” *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 381–387, 2003.

[98] P. Stephenson, I. Lungu, M. Paun, I. Silvas, and G. Tupu, “Tariff development for consumer groups in internal european electricity markets,” in *16th International Conference and Exhibition on Electricity Distribution, 2001. Part 1: Contributions. CIRED.(IEE Conf. Publ No. 482)*, vol. 5, pp. 5–pp, IET, 2001.

[99] C. Chen, M. Kang, J. Hwang, and C. Huang, “Synthesis of power system load profiles by class load study,” *International Journal of Electrical Power & Energy Systems*, vol. 22, no. 5, pp. 325–330, 2000.

[100] D. Wang, W. Xie, J. Pei, and Z. Lu, “Moving area detection based on estimation of static background,” *J Inform Comput Sci*, vol. 2, no. 1, pp. 129–134, 2005.

[101] M. B. Palacios and M. F. J. Steel, “Non-gaussian bayesian geostatistical modeling,” *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 604–618, 2006.

[102] M. Azam and N. Bouguila, “Bounded generalized gaussian mixture model with ica,” *Neural Processing Letters*, vol. 49, no. 3, pp. 1299–1320, 2019.

[103] Z. Xian, M. Azam, and N. Bouguila, "Statistical modeling using bounded asymmetric gaussian mixtures: Application to human action and gender recognition," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 41–48, IEEE, 2021.

[104] S. J. Raudys, A. K. Jain, *et al.*, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[105] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[106] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[107] R. Caruana and D. Freitag, "Greedy attribute selection," in *Machine Learning Proceedings 1994*, pp. 28–36, Elsevier, 1994.

[108] "High-performance computing facility: Speed." https://www.concordia.ca/ginacody/aits/speed.html, 2018. [Online; accessed 05-April-2023].

[109] A. Rafati, H. R. Shaker, and S. Ghahghahzadeh, "Fault detection and efficiency assessment for hvac systems using non-intrusive load monitoring: A review," *Energies*, vol. 15, no. 1, p. 341, 2022.

[110] M. Rodr'ıguez Fern'andez, A. Cort'es Garc'ıa, I. Gonz'alez Alonso, and E. Zalama Casanova, "Using the big data generated by the smart home to improve energy efficiency management," *Energy Efficiency*, vol. 9, no. 1, pp. 249–260, 2016.

[111] X. Liu and P. S. Nielsen, "A hybrid ict-solution for smart meter data analytics," *Energy*, vol. 115, pp. 1710–1722, 2016.

[112] N. Al Khafaf, M. Jalili, and P. Sokolowski, "A novel clustering index to find optimal clusters size with application to segmentation of energy consumers," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 346–355, 2020.

[113] G. Chicco, R. Napoli, and F. Piglione, “Comparisons among clustering techniques for electricity customer classification,” *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933–940, 2006.

[114] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” *Energy*, vol. 42, no. 1, pp. 68–80, 2012.

[115] R. Li, F. Li, and N. D. Smith, “Multi-resolution load profile clustering for smart metering data,” *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4473–4482, 2016.

[116] S. V. Verdú, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, “Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps,” *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1672–1682, 2006.

[117] G. Coke and M. Tsao, “Random effects mixture models for clustering electrical load series,” *Journal of time series analysis*, vol. 31, no. 6, pp. 451–464, 2010.

[118] F. McLoughlin, A. Duffy, and M. Conlon, “Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study,” *Energy and buildings*, vol. 48, pp. 240–248, 2012.

[119] N. Bouguila and W. Fan, *Mixture models and applications*. Springer, 2020.

[120] D. Peel and G. J. McLachlan, “Robust mixture modelling using the t distribution,” *Statistics and computing*, vol. 10, no. 4, pp. 339–348, 2000.

[121] C. Liu and D. B. Rubin, “Ml estimation of the t distribution using em and its extensions, ecm and ecme,” *Statistica Sinica*, pp. 19–39, 1995.

[122] X. Wei and Z. Yang, “The infinite student’s t-factor mixture analyzer for robust clustering and classification,” *Pattern Recognition*, vol. 45, no. 12, pp. 4346–4357, 2012.

[123] T. Elguebaly and N. Bouguila, “A nonparametric bayesian approach for enhanced pedestrian detection and foreground segmentation,” in *CVPR 2011 WORKSHOPS*, pp. 21–26, IEEE, 2011.

[124] Z. Gao, B. Belzer, and J. Villasenor, “A comparison of the z, e/sub 8/, and leech lattices for quantization of low-shape-parameter generalized gaussian sources,” *IEEE Signal Processing Letters*, vol. 2, no. 10, pp. 197–199, 1995.

[125] T. Elguebaly and N. Bouguila, “Finite asymmetric generalized gaussian mixture models learning for infrared object detection,” *Computer Vision and Image Understanding*, vol. 117, no. 12, pp. 1659–1671, 2013.

[126] A. Hyvärinen and P. Hoyer, “Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces,” *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[127] S. Bedingfield, D. Alahakoon, H. Genegedera, and N. Chilamkurti, “Multi-granular electricity consumer load profiling for smart homes using a scalable big data algorithm,” *Sustainable Cities and Society*, vol. 40, pp. 611–624, 2018.

[128] Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, “Sparse and redundant representation-based smart meter data compression and pattern extraction,” *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2142–2151, 2016.

[129] J. Yang and V. Honavar, “Feature subset selection using a genetic algorithm,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44–49, 1998.

[130] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, “Feature construction and calibration for clustering daily load curves from smart-meter data,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 645–654, 2016.

[131] F. Iglesias, T. Zseby, and A. Zimek, “Absolute cluster validity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2096–2112, 2019.

[132] F. N. Melzi, A. Same, M. H. Zayani, and L. Oukhellou, “A dedicated mixture model for clustering smart meter data: identification and analysis of electricity consumption behaviors,” *Energies*, vol. 10, no. 10, p. 1446, 2017.

[133] H. D.-G. Acquah, “Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship,” *Journal of Development and Agricultural Economics*, vol. 2, no. 1, pp. 001–006, 2010.

[134] C. S. Wallace and D. L. Dowe, “Mml clustering of multi-state, poisson, von mises circular and gaussian distributions,” *Statistics and Computing*, vol. 10, no. 1, pp. 73–83, 2000.

[135] C. S. Wallace and P. R. Freeman, “Estimation and inference by compact coding,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 49, no. 3, pp. 240–252, 1987.

[136] Y. Agusta and D. L. Dowe, “Unsupervised learning of gamma mixture models using minimum message length,” in *Proceedings of the 3rd IASTED Conference on Artificial Intelligence and Applications*, pp. 457–462, Acta Press Benalmadena, 2003.

[137] M. Azam and N. Bouguila, “Bounded laplace mixture model with applications to image clustering and content based image retrieval,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 558–563, IEEE, December 2018.

[138] T. Elguebaly and N. Bouguila, “Background subtraction using finite mixtures of asymmetric gaussian distributions and shadow detection,” *Machine vision and applications*, vol. 25, no. 5, pp. 1145–1162, 2014.

[139] C. S. Wallace and D. M. Boulton, “An information measure for classification,” *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[140] M. A. T. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[141] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*, vol. 290. Springer Science & Business Media, 2013.

[142] N. Bouguila and D. Ziou, “High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 10, pp. 1716–1731, 2007.

[143] J. C. Bezdek and N. R. Pal, “Some new indexes of cluster validity,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 301–315, 1998.

[144] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[145] T. Cali’nski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[146] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

[147] M. S. Allili, “Wavelet modeling using finite mixtures of generalized gaussian distributions: application to texture discrimination and retrieval,” *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1452–1464, 2011.

[148] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: an overview,” *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[149] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, pp. 461–464, 1978.

[150] M. Azam, B. Alghabashi, and N. Bouguila, “Multivariate bounded asymmetric gaussian mixture model,” *Mixture Models and Applications*, pp. 61–80, 2020.

[151] A. Boulmerka and M. Allili, “Thresholding-based segmentation revisited using mixtures of generalized Gaussian distributions,” *Proceedings Of The 21st International Conference On Pattern Recognition (ICPR2012)*, pp. 2894–2897, 2012.

[152] M. Allili and N. Baaziz, “Contourlet-based texture retrieval using a mixture of generalized Gaussian distributions,” *International Conference On Computer Analysis Of Images And Patterns*, pp. 446–454, 2011.

[153] M. Allili, D. Ziou, N. Bouguila, and S. Boutemedjet, “Image and video segmentation by combining unsupervised generalized Gaussian mixture modeling and feature selection,” *IEEE Transactions On Circuits And Systems For Video Technology*, vol. 20, pp. 1373–1377, 2010.

[154] J. Lee and A. Nandi, “Parameter estimation of the asymmetric generalised gaussian family of distributions,” *IEE Colloquium On Statistical Signal Processing (Ref. No. 1999/002)*, pp. 9–1, 1999.

[155] N. Nacereddine, S. Tabbone, D. Ziou, and L. Hamami, “Asymmetric generalized Gaussian mixture models and em algorithm for image segmentation,” *2010 20th International Conference On Pattern Recognition*, pp. 4557–4560, 2010.

[156] A. Ayebo and T. Kozubowski, “An asymmetric generalization of Gaussian and laplace laws,” *Journal Of Probability And Statistical Science*, vol. 1, pp. 187–210, 2003.

[157] N. Bouguila and D. A. Ziou, “Dirichlet process mixture of generalized dirichlet distributions for proportional data modeling,” *IEEE Transactions On Neural Networks*, vol. 21, pp. 107–122, 2009.

[158] H. A. Akaike, “new look at the statistical model identification,” *IEEE Transactions On Automatic Control*, vol. 19, pp. 716–723, 1974.

[159] T. Rossi and J. A. s. Braun, “rule-based fault detection and diagnostic method for vapor compression air conditioners,” *HvacR Research*, vol. 3, pp. 19–37, 1997.

[160] A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, and K. H. Johansson, “Estimation of building occupancy levels through environmental signals deconvolution,” in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pp. 1–8, 2013.

[161] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez, “An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network,” *Energy and Buildings*, vol. 42, no. 7, pp. 1038–1046, 2010.

[162] M. Azam, M. Blayo, J.-S. Venne, and M. Allegue-Martinez, “Occupancy estimation using wifi motion detection via supervised machine learning algorithms,” in *2019 ieee global conference on signal and information processing (GlobalSIP)*, pp. 1–5, IEEE, 2019.

[163] R. P. Browne, P. D. McNicholas, and M. D. Sparling, “Model-based learning using a mixture of mixtures of gaussian and uniform distributions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 814–817, 2011.

[164] K. Bruton, P. Raftery, P. O’Donovan, N. Aughney, M. Keane, and D. O’Sullivan, “Development and alpha testing of a cloud based automated fault detection and diagnosis tool for air handling units,” *Automation In Construction*, vol. 39, pp. 70–83, 2014.

[165] M. Mirnaghi and F. Haghighat, “Fault detection and diagnosis of large-scale hvac systems in buildings using data-driven methods: A comprehensive review,” *Energy And Buildings*, vol. 229, 2020.

[166] H. Yang, T. Zhang, H. Li, D. Woradechjumroen, and X. H. e. Liu, “unitary: fault detection and diagnosis,” *Encyclopedia Of Energy Engineering And Technology, Second Edition*, pp. 854–864, 2014.

[167] B. Li, F. Cheng, X. Zhang, C. Cui, and W. A. Cai, “novel semi-supervised data-driven method for chiller fault diagnosis with unlabeled data,” *Applied Energy*, vol. 285, 2021.

[168] K. Yan, A. Chong, and Y. Mo, “Generative adversarial network for fault detection diagnosis of chillers,” *Building and Environment*, vol. 172, p. 106698, 2020.

[169] G. Kaler Jr, “Expert system predicts service,” *Heating, piping and air conditioning*, vol. 60, no. 11, pp. 99–101, 1988.

[170] T. Reddy and K. Andersen, “An evaluation of classical steady-state off-line linear parameter estimation methods applied to chiller performance data,” *HvacR Research*, vol. 8, pp. 101–124, 2002.

[171] G. Xu, *HVAC system study: a data-driven approach*. PhD thesis, The University of Iowa, 2012.

[172] S. Mishra, A. Glaws, D. Cutler, S. Frank, M. Azam, F. Mohammadi, and J.-S. Venne, "Unified architecture for data-driven metadata tagging of building automation systems," *Automation in Construction*, vol. 120, p. 103411, 2020.

[173] R. Yang and G. Rizzoni, "Comparison of model-based vs. data-driven methods for fault detection and isolation in engine idle speed control system," in *Annual Conference of the PHM Society*, vol. 8, 2016.

[174] R. Yang, K. Li, E. Lemarchand, and T. Fen-Chong, "Micromechanics modeling the solute diffusivity of unsaturated granular materials," *International Journal Of Multiphase Flow*, vol. 79, pp. 1–9, 2016.

[175] F. Mtibaa, K.-K. Nguyen, M. Azam, A. Papachristou, J.-S. Venne, and M. Cheriet, "Lstm-based indoor air temperature prediction framework for hvac systems in smart buildings," *Neural Computing and Applications*, vol. 32, pp. 17569–17585, 2020.

[176] M. C. Comstock, J. E. Braun, and E. A. Groll, "A survey of common faults for chillers/discussion," *ASHRAE Transactions*, vol. 108, pp. 819–827, 2002.

[177] G. Li and Y. Hu, "An enhanced pca-based chiller sensor fault detection method using ensemble empirical mode decomposition based denoising," *Energy And Buildings*, vol. 183, pp. 311–324, 2019.

[178] M. Amayri, A. Arora, S. Ploix, S. Bandhyopadyay, Q. Ngo, and V. Badarla, "Estimating occupancy in heterogeneous sensor environment," *Energy And Buildings*, vol. 129, pp. 46–58, 2016.

[179] C. Roulet, P. Cretton, R. Fritsch, and J. Scartezzini, "Stochastic model of inhabitant behavior with regard to ventilation," in *Proceedings of the 12th AIVC Conference, Ottawa*, 1991.

[180] J. Page, D. Robinson, and J. Scartezzini, "Stochastic simulation of occupant presence and behaviour in buildings," in *Proc. Tenth Int. IBPSA Conf: Building Simulation*, pp. 757–764, 2007.

[181] F. Haldi and D. Robinson, “Interactions with window openings by office occupants,” *Building And Environment*, vol. 44, pp. 2378–2395, 2009.

[182] A. Kashif, J. Dugdale, and S. Ploix, “Simulating occupants’behavior for energy waste reduction in dwellings: A multiagent methodology,” *Advances in Complex Systems*, vol. 16, no. 04n05, p. 1350022, 2013.

[183] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, “Occupancy-driven energy management for smart building automation,” *Proceedings Of The 2nd ACM Workshop On Embedded Sensing Systems For Energy-efficiency In Building*, pp. 1–6, 2010.

[184] V. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A. Cerpa, M. Sohn, and S. Narayanan, “Energy efficient building environment control strategies using real-time occupancy measurements,” *Proceedings Of The First ACM Workshop On Embedded Sensing Systems For Energy-efficiency In Buildings*, pp. 19–24, 2009.

[185] A. Kamthe, V. Erickson, M. Carreira-Perpiñán, and A. Cerpa, “Enabling building energy auditing using adapted occupancy models,” in *Proceedings of the third ACM workshop on Embedded sensing systems for energy-efficiency in buildings*, pp. 31–36, 2011.

[186] C. Martani, D. Lee, P. Robinson, R. Britter, and C. E. Ratti, “Studying the dynamic relationship between building occupancy and energy consumption,” *Energy And Buildings*, vol. 47, pp. 584–591, 2012.

[187] M. Ebling and M. Corner, “It’s all about power and those pesky power vampires,” *IEEE Pervasive Computing*, vol. 8, pp. 12–13, 2009.

[188] V. Erickson and A. Cerpa, “Occupancy based demand response hvac control strategy,” *Proceedings Of The 2nd ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Building*, pp. 7–12, 2010.

[189] D. Delaney, G. O’Hare, and A. Ruzzelli, “Evaluation of energy-efficiency in lighting systems using sensor networks,” *Proceedings Of The First ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings*, pp. 61–66, 2009.

[190] A. Kamthe, L. Jiang, M. Dudys, and A. Cerpa, “Scopes: Smart cameras object position estimation system,” in *Wireless Sensor Networks: 6th European Conference, EWSN 2009, Cork, Ireland, February 11-13, 2009. Proceedings 6*, pp. 279–295, Springer, 2009.

[191] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. T. s. t. Whitehouse, “using occupancy sensors to save energy in homes,” *Proceedings Of The 8th ACM Conference On Embedded Networked Sensor Systems*, pp. 211–224, 2010.

[192] S. Wang and X. Jin, “Co2-based occupancy detection for on-line outdoor air flow control,” *Indoor And Built Environment*, vol. 7, pp. 165–181, 1998.

[193] M. Amayri and S. Ploix, “Decision tree and parametrized classifier for estimating occupancy in energy management,” *2018 5th International Conference On Control, Decision And Information Technologies (CoDIT)*, pp. 397–402, 2018.

[194] P. Dongre, A. Aldrees, and D. Gracanin, “Clustering appliance energy consumption data for occupant energy-behavior modeling,” *Proceedings Of The 8th ACM International Conference On Systems For Energy-Efficient Buildings, Cities, And Transportation*, pp. 290–293, 2021.

[195] D. Murray, J. Liao, L. Stankovic, V. Stankovic, R. Hauxwell-Baldwin, C. Wilson, M. Coleman, T. Kane, and S. Firth, “A data management platform for personalised real-time energy feedback,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pp. 1289–1297, ACM, 2015.

[196] B. G. Leroux and M. L. Puterman, “Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models,” *Biometrics*, pp. 545–558, 1992.

[197] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[198] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, “A review of human activity recognition methods,” *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[199] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng, "Transferable feature representation for visible-to-infrared cross-dataset human action recognition," *Complexity*, vol. 2018, pp. 1–20, 2018.

[200] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[201] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 32–36, IEEE, 2004.

[202] M. Fu, N. Chen, Z. Huang, K. Ni, Y. Liu, S. Sun, and X. Ma, "Human action recognition: A survey," in *Signal and Information Processing, Networking and Computers: Proceedings of the 5th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*, pp. 69–77, Springer, 2019.

[203] Z. Moghaddam and M. Piccardi, "Training initialization of hidden markov models in human action recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 394–408, 2013.

[204] C. Gao, Y. Du, J. Liu, L. Yang, and D. Meng, "A new dataset and evaluation for infrared action recognition," in *Computer Vision: CCF Chinese Conference, CCCV 2015, Xi'an, China, September 18-20, 2015, Proceedings, Part II*, pp. 302–312, Springer, 2015.

[205] E. Epaillard and N. Bouguila, "Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas," *Pattern Recognition*, vol. 55, pp. 125–136, 2016.

[206] E. Epaillard and N. Bouguila, "Hidden markov models based on generalized dirichlet mixtures for proportional data modeling," in *Artificial Neural Networks in Pattern Recognition: 6th IAPR TC 3 International Workshop, ANNPR 2014, Montreal, QC, Canada, October 6-8, 2014. Proceedings 6*, pp. 71–82, Springer, 2014.

[207] E. Epaillard and N. Bouguila, “Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 4, pp. 1034–1047, 2018.

[208] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou, “Robust sequential data modeling using an outlier tolerant hidden markov model,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 9, pp. 1657–1669, 2008.

[209] S. Ali and N. Bouguila, “Variational learning of beta-liouville hidden markov models for infrared action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

[210] Z. Xian, M. Azam, M. Amayri, W. Fan, and N. Bouguila, “Bounded asymmetric gaussian mixture-based hidden markov models,” in *Hidden Markov Models and Applications*, pp. 33–58, Springer, 2022.

[211] N. Bouguila, W. Fan, and M. Amayri, *Hidden Markov Models and Applications*. Springer, 2022.

[212] S. Adams, P. A. Beling, and R. Cogill, “Feature selection for hidden markov models and hidden semi-markov models,” *IEEE Access*, vol. 4, pp. 1642–1657, 2016.

[213] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-h. Zahzah, “Human pose estimation from monocular images: A comprehensive survey,” *Sensors*, vol. 16, no. 12, p. 1966, 2016.

[214] M. Raptis, M. Bustreo, and S. Soatto, “Time warping under dynamic constraints,” in *Eleventh IEEE International Conference on Computer Vision, Workshop on Dynamical Vision*, (Rio de Janeiro, Brazil), IEEE, 2007.

[215] M. Mackay, R. G. Fenton, and B. Benhabib, “A real-time visual action-recognition framework for time-varying-geometry objects,” in *2010 IEEE International Conference on Automation Science and Engineering*, pp. 922–927, IEEE, 2010.

[216] R. Vezzani, D. Baltieri, and R. Cucchiara, “Hmm based action recognition with projection histogram features,” in *Recognizing Patterns in Signals, Speech, Images and Videos: ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports*, pp. 286–293, Springer, 2010.

[217] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. A. Velastin, “Recognizing human actions using silhouette-based hmm,” in *2009 Sixth IEEE international conference on advanced video and signal based surveillance*, pp. 43–48, IEEE, 2009.

[218] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, “Detecting human behavior models from multimodal observation in a smart home,” *IEEE Transactions on automation science and engineering*, vol. 6, no. 4, pp. 588–597, 2008.

[219] N. Li and D. Xu, “Action recognition using weighted three-state hidden markov model,” in *2008 9th International Conference on Signal Processing*, pp. 1428–1431, IEEE, 2008.

[220] M. Ahmad and S.-W. Lee, “Hmm-based human action recognition using multiview image sequences,” in *18th international conference on pattern recognition (ICPR’06)*, vol. 1, pp. 263–266, IEEE, 2006.

[221] R. Poppe, “A survey on vision-based human action recognition,” *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[222] L. Wang and D. Suter, “Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[223] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pp. 65–72, IEEE, 2005.

[224] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[225] N. Ikizler and D. A. Forsyth, “Searching for complex human activities with no visual examples,” *International Journal of Computer Vision*, vol. 80, no. 3, pp. 337–357, 2008.

[226] J. Kittler, P. Pudil, and P. Somol, “Advances in statistical feature selection,” in *Advances in Pattern Recognition—ICAPR 2001: Second International Conference Rio de Janeiro, Brazil, March 11–14, 2001 Proceedings*, pp. 427–436, Springer, 2001.

[227] C. Lee and G. G. Lee, “Information gain and divergence-based feature selection for machine learning-based text categorization,” *Information processing & management*, vol. 42, no. 1, pp. 155–165, 2006.

[228] H. Uğuz, “A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm,” *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[229] M. Graña, M. Termenon, A. Savio, A. Gonzalez-Pinto, J. Echeveste, J. Pérez, and A. Besga, “Computer aided diagnosis system for alzheimer disease using brain diffusion tensor imaging features selected by pearson’s correlation,” *Neuroscience letters*, vol. 502, no. 3, pp. 225–229, 2011.

[230] N. A. Khan, S. A. Waheeb, A. Riaz, and X. Shang, “A three-stage teacher, student neural networks and sequential feed forward selection-based feature selection approach for the classification of autism spectrum disorder,” *Brain sciences*, vol. 10, no. 10, p. 754, 2020.

[231] H. Eom, Y. Son, and S. Choi, “Feature-selective ensemble learning-based long-term regional pv generation forecasting,” *IEEE access*, vol. 8, pp. 54620–54630, 2020.

[232] C. P. Ezenkwu, U. I. Akpan, and B. U.-A. Stephen, “A class-specific metaheuristic technique for explainable relevant feature selection,” *Machine Learning with Applications*, vol. 6, p. 100142, 2021.

[233] S. Maldonado and R. Weber, “A wrapper method for feature selection using support vector machines,” *Information Sciences*, vol. 179, no. 13, pp. 2208–2217, 2009.

[234] W. Zhang and Z. Yin, "Eeg feature selection for emotion recognition based on cross-subject recursive feature elimination," in *2020 39th Chinese Control Conference (CCC)*, pp. 6256–6261, IEEE, 2020.

[235] J. H. Holland, "Genetic algorithms," *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.

[236] R. Nasfi, M. Amayri, and N. Bouguila, "A novel approach for modeling positive vectors with inverted dirichlet-based hidden markov models," *Knowledge-Based Systems*, vol. 192, p. 105335, 2020.

[237] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden markov models," *IEEE transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[238] J. Zhou and X.-P. Zhang, "An ica mixture hidden markov model for video content analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1576–1586, 2008.

[239] W. Fan, H. Sallay, N. Bouguila, and S. Bourouis, "A hierarchical dirichlet process mixture of generalized dirichlet distributions for feature selection," *Computers & Electrical Engineering*, vol. 43, pp. 48–65, 2015.

[240] J. Montero and L. Sucar, "Feature selection for visual gesture recognition using hidden markov models," in *Proceedings of the Fifth Mexican International Conference in Computer Science, 2004. ENC 2004.*, pp. 196–203, IEEE, 2004.

[241] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, vol. 25, no. 7, pp. 767–775, 2004.

[242] S. Adams, *Simultaneous feature selection and parameter estimation for hidden Markov models*. PhD thesis, University of Virginia, Charlottesville, VA, USA, 2015.

[243] M. Azam and N. Bouguila, “Speaker verification using adapted bounded gaussian mixture model,” in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 300–307, IEEE, 2018.

[244] M. Azam and N. Bouguila, “Multivariate bounded support asymmetric generalized gaussian mixture model with model selection using minimum message length,” *Expert Systems with Applications*, vol. 204, p. 117516, 2022.

[245] M. Bicego, U. Castellani, and V. Murino, “A hidden markov model approach for appearance-based 3d object recognition,” *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2588–2599, 2005.

[246] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, “An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition,” *Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[247] E. L. Andrade, S. Blunsden, and R. B. Fisher, “Hidden markov models for optical flow analysis in crowds,” in *18th international conference on pattern recognition (ICPR’06)*, vol. 1, pp. 460–463, IEEE, 2006.

[248] J. Sepúlveda and S. A. Velastin, “F1 score assesment of gaussian mixture background subtraction algorithms using the muhavi dataset,” 2015.

[249] R. Shi, K. N. Ngan, and S. Li, “Jaccard index compensation for object segmentation evaluation,” in *2014 IEEE international conference on image processing (ICIP)*, pp. 4457–4461, IEEE, 2014.

[250] M. D. Katzman, “Analyzing a portion of the roc curve,” *Med. Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[251] A. Rosenberg and J. Hirschberg, “V-measure: A conditional entropy-based external cluster evaluation measure,” in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

[252] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[253] A. Amelio and C. Pizzuti, "Correction for closeness: Adjusting normalized mutual information measure for clustering comparison," *Computational Intelligence*, vol. 33, no. 3, pp. 579–601, 2017.

[254] P. Bagade, A. Banerjee, and S. K. Gupta, "Optimal design for symbiotic wearable wireless sensors," in *2014 11th International Conference on Wearable and Implantable Body Sensor Networks*, pp. 132–137, IEEE, 2014.

[255] W. Y. Toh, Y. K. Tan, W. S. Koh, and L. Siek, "Autonomous wearable sensor nodes with flexible energy harvesting," *IEEE sensors journal*, vol. 14, no. 7, pp. 2299–2306, 2014.

[256] A. Godfrey, "Wearables for independent living in older adults: Gait and falls," *Maturitas*, vol. 100, pp. 16–26, 2017.

[257] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2014.

[258] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pp. 138–143, IEEE, 2009.

[259] O. A. Atoyebi, A. Stewart, and J. Sampson, "Use of information technology for falls detection and prevention in the elderly," *Ageing international*, vol. 40, pp. 277–299, 2015.

[260] L. Gao, A. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," *Medical engineering & physics*, vol. 36, no. 6, pp. 779–785, 2014.

[261] K. Altun and B. Barshan, "Human activity recognition using inertial/magnetic sensor units," in *Human Behavior Understanding: First International Workshop, HBU 2010, Istanbul, Turkey, August 22, 2010. Proceedings 1*, pp. 38–51, Springer, 2010.

[262] L. Malott, P. Bharti, N. Hilbert, G. Gopalakrishna, and S. Chellappan, "Detecting self-harming activities with wearable devices," in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 597–602, IEEE, 2015.

[263] P. Bharti, A. Panwar, G. Gopalakrishna, and S. Chellappan, "Watch-dog: Detecting self-harming activities from wrist worn accelerometers," *IEEE journal of biomedical and health informatics*, vol. 22, no. 3, pp. 686–696, 2017.

[264] E. K. Choe, J. A. Kientz, S. Halko, A. Fonville, D. Sakaguchi, and N. F. Watson, "Opportunities for computing to support healthy sleep behavior," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pp. 3661–3666, 2010.

[265] J. S. Bauer, S. Consolvo, B. Greenstein, J. Schooler, E. Wu, N. F. Watson, and J. Kientz, "Shuteye: encouraging awareness of healthy sleep recommendations with a mobile, peripheral display," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1401–1410, 2012.

[266] N. Oliver and F. Flores-Mangas, "Healthgear: Automatic sleep apnea detection and monitoring with a mobile phone," *Journal of Communications*, vol. 2, no. 2, pp. 1–9, 2007.

[267] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 477–486, 2014.

[268] M. Kay, E. K. Choe, J. Shepherd, B. Greenstein, N. Watson, S. Consolvo, and J. A. Kientz, "Lullaby: a capture & access system for understanding the sleep environment," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 226–234, 2012.

[269] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*, pp. 108–109, IEEE, 2012.

[270] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–8, 2012.

[271] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[272] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033–2042, 2013.

[273] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14363–14373, 2021.

[274] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2990–3001, 2020.

[275] E. A. Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks," *Sensors*, vol. 13, no. 7, pp. 8750–8770, 2013.

[276] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[277] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, (Long Beach, CA, USA), pp. 7272–7282, Neural Information Processing Systems Foundation, December 2017.

[278] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of computer vision*, vol. 37, no. 2, pp. 151–172, 2000.

[279] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[280] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, pp. 1395–1402, IEEE, 2005.

[281] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, pp. 299–318, 2008.

[282] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human action detection by boosting efficient motion features," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 522–529, IEEE, 2009.

[283] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *2007 IEEE 11th international conference on computer vision*, pp. 1–8, Ieee, 2007.

[284] S. Rajaraman, S. K. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. R. Thoma, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, 2018.

[285] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2061–2068, IEEE, 2010.

[286] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*, pp. 140–153, Springer, 2010.

[287] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[288] K. Schindler and L. Van Gool, “Action snippets: How many frames does human action recognition require?,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.