Investigate Organizational Member Engagement Through Financial X-ray and Artificial Neural Networks

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

by

David Hason Rudd
Under the supervision of Professor Guandong Xu and Dr. Huan Huo

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

November 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *David HASON RUDD*, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy Analytics, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research was supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

SIGNATURE: ________________________________

[David HASON RUDD]

DATE: 25th November, 2023

PLACE: Sydney, Australia

ABSTRACT

Abstract. Comprehensive understanding of member engagement and churn is imperative for employees within financial institutions and associations, necessitating shifting from conventional approaches toward more sophisticated analytical paradigms. Employing customer voice (CV), financial literacy (FL), and customer relationship management (CRM) data for churn analysis helps to define member engagement level, facilitating effective retention strategies and fostering long-term loyalty for sustained growth. Members with improved financial knowledge are better equipped to make advantageous decisions and less likely to churn due to misconceptions or unmet expectations. Concurrently, in an era where Telephonic interactions have become the norm post-COVID-19, and emotional content derived from such conversational interactions currently provides real-time insights into member's sentiments, and can be utilized as a predictor for churn modeling. Although many previous studies have explored helpful information to analyze member's behavior for churn, they often overlooked bridging member engagement and churn through a holistic view of members' interactions, emotions, FL, and CRM data. Several approaches to addressing these issues have been introduced using single data sources, e.g., transactional, demographic, and textual data, which are not multifaceted views of member behavior. Current efforts are limited to three main challenges. First, transactional data employed in several recent studies only reflected prediction outcomes, rather than experience or underlying causes for churn. Second, although demographic data have been employed in many studies; static data does not capture dynamic customer satisfaction. Third, social media data (textual data) has been employed in a few previous studies, but textual input is noisy and lacks personalized insights, such as voice interaction and financial skills. Therefore, this thesis leverages a multimodal modeling approach to capture multifaceted insights for member engagement.

The main themes of this thesis include

1. Introduce novel speech emotion recognition (SER) methods, developing a VGG-optiVMD algorithm to capture real-time emotions from CV data, enabling early detection of dissatisfaction and personalized interactions. This approach leverages advanced acoustic analysis to improve customer service responsiveness and personalize interaction strategies, directly impacting customer retention.

2. Develop an SER model using CV signal processing, harmonic and percussive components from the Mel Spectrogram acoustic feature, and CNN-VGG16 architectures.

This model enhances the accuracy of emotion detection, providing deeper insights into customer sentiments and enabling more effective communication strategies.

3. Develop SMOGN-COREG semi-supervised machine-learning techniques to extract patterns from unlabeled financial network data and subsequently predict FL levels. This technique maximizes the use of available data, reducing the need for extensive labeled datasets and lowering the barriers to comprehensive financial literacy analysis.

4. Develop a causal model to understand root causes for churn in member-centric organizations. This model helps tailor interventions to prevent churn and improve customer engagement by identifying the underlying factors contributing to customer departure.

5. Develop a multimodal hybrid fusion learning model that not only integrates FL metrics, behavioral indicators, and voice emotional features; but also incorporates essential member engagement aspect, significantly enhancing churn prediction precision. This holistic approach provides a nuanced understanding of churn, enabling precise targeting of at-risk customers based on a comprehensive data profile.

6. Develop a state-of-the-art model by applying multifaceted neural network architectures, data augmentation strategies, and emotion recognition algorithms. These techniques advance the model's learning capabilities, ensuring robust performance even in complex and dynamic data environments.

The present study is the first to propose a multimodal hybrid fusion technique effectively combining CV, FL, and CRM data, and hence providing deeper understanding of member engagement for churn risk analysis. Empirical results from this thesis demonstrate the developed methods' advantages and effectiveness, which will be valuable CRM research. This study proposes a comprehensive framework for organizations to enhance member engagement and minimize churn by integrating disparate but interrelated threads including financial skill, member sentiment, and financial behavioral data. The proposed framework provides a strategic blueprint for organizations to ensure sustainable growth and build lasting relationships with their members.

DEDICATION

To my loved ones ...

ACKNOWLEDGMENTS

Studying for the Doctorate of Philosophy is a long journey, and this thesis would not have been possible without the support and encouragement from my supervisors, friends, and relatives. A difficult part of writing the thesis has been how to adequately express my sincere gratitude.

My sincerest gratitude to my supervisors, Professor Guandong XU and Doctor Huan HUO, for their continuous support, motivation, enthusiasm, and inspiration and their invaluable advice and discussion during our weekly meetings. Their guidance will remain a positive effect throughout my future career path. I also thank them for their valuable comments and discussion regarding the various research manuscripts that arose along my Ph.D. journey. My sincere gratitude for the comprehensive financial assistance provided by the Australian Government Research Training Program Stipend (RTP) and the Faculty of Engineering and Information Technology (FEIT) Scholarship throughout my Ph.D. study.

A large portion of this thesis has been published in various journals after peer review by anonymous reviewers. I extend special thanks to those reviewers who provided valuable suggestions to improve the research papers. My thanks to the friends who accompanied me throughout my Ph.D. journey, the many conversations and fun we experienced will remain unforgettable. Finally, I express my love and boundless gratitude to my parents for their unconditional support and sacrifice. Their emotional and personal support has been critical to allowing me to completely concentrate on the research and has significantly impacted on the final thesis quality.

David HASON RUDD
Sydney, Australia, 2023.

List of Publications

Conferences

1. **David Hason Rudd**, Huo, H., Xu, G. (2022). Predicting Financial Literacy via Semi-supervised Learning. In: Long, G., Yu, X., Wang, S. (eds) AI 2021: Advances in Artificial Intelligence. AI 2022. Lecture Notes in Computer Science(), vol 13151. pp. 304-319, Springer, Cham. https://doi.org/10.1007/978-3-030-97546-3_25. **(Refer to Chapter 4)**

2. **David Hason Rudd**, H. Huo and G. Xu, "Causal Analysis of Customer Churn Using Deep Learning," 2021 International Conference on Digital Society and Intelligent Systems (DSInS), Chengdu, China, 2021, pp. 319-324, **(Refer to Chapter 5)**

3. **David Hason Rudd**, Rudd, D.H., Huo, H., Xu, G. (2022). Leveraged Mel Spectrograms Using Harmonic and Percussive Components in Speech Emotion Recognition. In: Gama, J., Li, T., Yu, Y., Chen, E., Zheng, Y., Teng, F. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2022. Lecture Notes in Computer Science(), vol 13281. pp. 392-404 Springer, Cham. https://doi.org/10.1007/978-3-031-05936-0_31. **(Refer to Chapter 3)**

4. **David Hason Rudd**, Huo, H., Xu, G. (2023). An Extended Variational Mode Decomposition Algorithm Developed Speech Emotion Recognition Performance. In: Kashima, H., Ide, T., Peng, WC. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2023. Lecture Notes in Computer Science(), vol 13937. pp. 291-331 Springer, Cham. https://doi.org/10.1007/978-3-031-33380-4_17 **(Refer to Chapter 3)**

5. **David Hason Rudd**, Churn Prediction via Multimodal Fusion Learning: Integrating Customer Financial Literacy, Voice, and Behavioral Data. BESC2023 **(Refer to Chapter 6)**

Journal

1. **David Hason Rudd**, Huo, H. & Xu, G. Improved Churn Causal Analysis Through Restrained High-Dimensional Feature Space Effects in Financial Institutions. Hum-Cent Intell Syst vol. 2, pp. 70-80 (2022), doi: 10.1007/s44230-022-00006-y. **(Refer to Chapter 5)**

Table of Contents

LIST OF FIGURES

FIGURE **Page**

List of Tables

Table **Page**

Chapter 1

Introduction

Member engagement is paramount to organizational vitality. Engaged members actively participate in events and fully utilize the services offered, sustaining the organization's core purpose. High engagement levels translate into numerous benefits for the organization, with reduced churn rates being a particular advantage, since organizational growth is intrinsically linked to its ability to retain existing members. Organizations develop an environment that encourages members to renew their commitment by providing dynamic and compelling content that resonates with them. This creates a stable and reliable member base that is indispensable for the organization's long-term success and sustainability. However, maintaining this engagement level becomes increasingly challenging as organizations navigate the complex global market dynamics. In the increasingly complex global market landscape, businesses constantly face dual challenges to fully understand member's engagement and retain them. Traditional customer relationship management (CRM) strategies evolve as industries grow and technology advances, with businesses now placing significant emphasis on both acquiring new members and ensuring existing member loyalty. Retaining members is not only crucial but also economically significant for sustained growth and profitability within the financial sector. For instance, American Express found that loyal customers are likely to spend up to 67% more than new ones [1]. To ensure a comprehensive analysis of organizational member engagement and churn, it's vital to address several identified gaps in the literature. Table 1.1 presents a summary of knowledge gaps. Traditional CRM strategies predominantly focus on member acquisition rather than retention, overlooking the cost-effectiveness of nurturing existing members. Current models often underutilize multimodal data, limiting the depth and accuracy of churn predictions. There's also a significant gap in integrating emotional data, which can preempt churn indications. Moreover, reliance on static demographic data fails to capture the dynamic nature of customer satisfaction, and most systems lack real-time prediction capabilities. These gaps highlight the necessity for enhanced multimodal integration and real-time analysis in churn prediction models.

Table 1.1: Summary of literature gaps and limitations

Gap	Description
1	Traditional CRM strategies emphasize member acquisition over retention, not recognizing the cost-effectiveness of nurturing existing members.
2	Current models often fail to utilize multimodal data fully, restricting the depth and accuracy of churn predictions.
3	Significant gap in integrating emotional data, which can provide early warnings of potential churn.
4	Over-reliance on static demographic data does not reflect the dynamic nature of customer satisfaction.
5	Most systems lack real-time prediction capabilities, limiting their effectiveness in proactive churn management.

This thesis investigates each facet of this multifaceted problem, providing insights, methodologies, and implications for businesses globally. Section 1.1 presents the research background and motivation, and Sections 1.2 to 1.4 discuss the objectives, limitations, and contributions, respectively. Section 1.5 describes the thesis structure.

1.1 Background and Motivation

1.1.1 Economic imperative to minimize churn rate

Historically, many businesses, particularly growing ones, have focused almost exclusively on member acquisition, using the simple logic that more members equaled more business. However, the current economic environment, marked by increased competition, has compelled businesses to reconsider this approach. Bringing in new members is a costly process, in terms of both monetary cost and resources expended. Retaining existing members, who have already been oriented into the business and its offerings, can be more cost-effective, providing larger transaction values and more frequent interactions [2].

This realization has led to development and widespread adoption of the churn rate metric to quantify the rate at which a business loses members, customers, or subscribers over a specified timeframe. High churn rate signals potential problems with product offerings or service quality and has direct financial implications due to associated lost revenue. Member-centric organizations must overcome high churn rate by addressing both the underlying causes of churn and reinvesting in acquiring new members to replace those who left [3].

1.1.2 Member churn and human behavior factors

Factors contributing to member churn are diverse, including voluntary churn, where members leave due to perceived better alternatives or dissatisfaction; involuntary churn, due to unavoidable circumstances, e.g., financial problems; incidental churn, triggered by relocation, etc.; and deliberate churn, rooted in unhappiness with service quality, pricing, or unmet expectations. In the information age, members are more aware and have a plethora of choices at their fingertips, hence the root causes for churn become more sophisticated, and understanding these become paramount.

Understanding human behavior, particularly in financial decision contexts, adds another layer of complexity. Cognitive biases are commonly pivotal in how individuals perceive and interact with financial services. For example biases, such as anchoring bias; availability heuristic; and bandwagon effects, where decisions are based on recent information, can significantly impact member trust in a financial institution, particularly in the aftermath of negative news or rumors [2, 4].

1.1.3 Leveraging emotion and financial literacy in churn prediction

The onset of the digital transformation in organizations and associated explosion in data availability and granularity have revolutionized churn prediction. Businesses are no longer limited to analyzing transactional data or demographic factors, and can now access diverse sources for member-related information.

One important churn prediction area is analyzing member emotions derived from voice and text interactions. Understanding these underlying sentiments allows businesses to predict potential member churn before it occurs. Financial literacy (FL) is another crucial factor, particularly within financial service organizations. Members' understanding of financial products and services is directly related to their satisfaction levels. A well-informed member can navigate financial products more efficiently, leading to reduced churn; whereas those who struggle might become gradually more dissatisfied, increasing the likelihood they move away from complex financial products [3]. A real-world example highlighting the link between customer emotion, financial literacy, and financial products is seen in the area of mortgage handling. When customers understand the implications of interest rates and the benefits of refinancing options, their satisfaction and trust in the financial institution increase. For instance, during the 2008 financial crisis, customers with higher financial literacy were less likely to default on mortgages because they were more likely to renegotiate their loan terms or refinance due to their understanding of changing interest rates. This knowledge not only impacts their financial decisions but also reduces frustration and anxiety, encouraging a stronger emotional bond with the financial service provider [5].

1.1.4 Multifaceted engagement analysis: diverse data for deeper insights

Sophisticated analytical methodologies are essential to understand the present plethora and diversity of member data. Multimodal machine learning appears to be a promising solution to this challenge [6], integrating multifaceted data modalities, ranging from conventional CRM repositories to sentiment analytics from member engagements, and even FL measures. Thus, multimodal learning endeavors to construct a detailed and holistic image of member behavior in financial organizations, creating a "financial X-Ray" for members. This expansive data perspective is pivotal to refine churn prediction algorithms. Combining diverse data sources ensures comprehensive understanding of member motives and actions, ensuring robust and strategically actionable churn predictions in real-world scenarios.

Member retention is a critical factor for business success in the current global market. Competition is intense, and digital advancements have led businesses to focus on member-centric models, where member churn, i.e., the rate at which customers discontinue services, indicates customer satisfaction and business health. Business models for finance, telecom, and e-commerce are strongly based on recurring subscriptions, hence churn can significantly affect revenue and harm brand reputation. Traditional approaches to managing churn have relied on transactional data and demographic analysis, but these methods have become inadequate due to the wealth of data now available, including text, images, and audio, which traditional methods may not fully utilize.

Member behavior is complex and influenced by cognitive, emotional, and situational factors. Cognitive biases, e.g., anchoring effects and availability heuristics, can significantly alter financial decision making and increase perceived risk, potentially leading to increased churn. Recent studies have highlighted the importance of distinguishing between immediate churn triggers and underlying churn predictors. Although triggers are direct causes for discontinuing service, predictors can offer subtle indicators for potential churn, and recognizing these predictors allows businesses to act before churn occurs.

To conduct a comprehensive analysis of churn and member engagement in financial organizations, it is essential to integrate various models that reflect member emotions, financial literacy, and churn propensity. Previous studies have primarily focused on utilizing a single model based on demographic and CRM data. However, there is a lack of research incorporating emotional and financial literacy features to predict member engagement and churn within organizations. While state-of-the-art single-model approaches exist, they are not necessarily tailored for analyzing churn and member engagement.

The current state-of-the-art in speech emotion recognition (SER) model is represented by the method proposed by Zhao et al. [7]. This method employs a combination of convolutional neural networks (CNN) and long short-term memory networks (LSTM) to analyze log Mel spectrograms for emotion recognition. It has achieved a notable accuracy of 95.89% on the EMODB database, outperforming various other methodologies listed in the study, making it a leading technique in the field of SER. In recent advancements within the field of financial literacy prediction, Rudd et al. [8] developed the SMOGN-COREG model, which leverages semi-supervised learning (SSL) techniques to enhance the accuracy of financial literacy predictions using unlabeled financial data. This innovative approach incorporates the Synthetic Minority Over-sampling Technique for Regression (SMOGN) alongside a co-regression (COREG) algorithm, effectively addressing the challenges posed by unbalanced datasets. The experimental outcomes demonstrated that the SMOGN-COREG model significantly outperforms traditional regression models, achieving higher prediction accuracy on several financial datasets. This underscores the potential of semi-supervised methods in utilizing unlabeled data to improve predictive performance in financial applications. In recent advancements within the field of financial literacy prediction. In the financial and banking sector, the latest state-of-the-art for customer churn prediction has been significantly advanced by the work of Tran et al. [9]. They explored the impact of customer segmentation using diverse machine learning models, including k-means clustering, k-nearest neighbors, logistic regression, decision trees, random forests, and support vector machines. Their findings indicated that the random forest model was most effective, achieving a 97% accuracy rate.

Effective churn management requires an advanced analytical framework that can handle multiple data types and provide a more profound understanding of customer behavior. Thus, there is a need for holistic churn analyses that incorporate diverse data types and ranges, including psychological and behavioral factors, to provide comprehensive explanations for member churn dynamics.

Given these realities, the motivation for my Ph.D. research is a compelling need, and I must recognize the challenges and potential within the modern data deluge. This study aims to navigate this duality by considering the following aspects.

1. **Bridge the evolving data landscape.** Member data is constantly, and significantly, growing. Hence, it is essential to harmonize various data streams into a unified, actionable narrative [10]. The proposed multimodal approach enables a more robust understanding of member behavior and churn patterns by combining data sources.

2. **Unbiased churn prediction.** Inherent biases in traditional churn prediction models, originating from over-reliance on singular data sources or overly simplistic algorithms, can distort reality. The proposed core approach will leverage advanced machine learning techniques on multiple member data sources to deliver more accurate and critically unbiased churn likelihood views and predictions.

3. **Financial X-Ray analytical approach to churn.** Transcending conventional method boundaries will enable holistic understanding of member churn. This 360-degree perspective, called the member Financial X-Ray, can explore deep into client behavior intricacies, offering insights to empower business' strategic and informed decisions. The Financial X-Ray is a comprehensive approach to understanding members' FL levels and behaviors by exploring various dimensions and factors contributing to churn, well beyond surface level analyses. Exploring the underlying dynamics and patterns that influence member decisions can identify key drivers and hidden trends for churn, and develop targeted strategies to mitigate member attrition. Businesses can leverage insights from Financial X-Ray to make confident data-driven decisions that align with member needs and preferences. This comprehensive understanding of member behavior will empower businesses

to optimize their offerings, enhance member engagement and experiences, and establish long-term relationships with clientele.

Contemporary business environment dynamism incorporating increasing data availability and evolving member behaviors require a paradigm shift in understanding and addressing member churn, empowering businesses to rethink traditional models. This thesis considers this challenging but promising interaction, aspiring to illuminate paths to greater member understanding, retention, and, ultimately, business success.

1.2 Research Objectives

Core objective. This thesis' main research objective is to develop a comprehensive multimodal framework leveraging diverse data sources, including member FL, behavioral data, and customer voice (CV) interactions. The specific output is to develop an actionable model that informs targeted engagement and retention strategies. Therefore, this thesis investigates intricate relationships among multiple modalities, including para-linguistic emotion detection from vocal interactions, member FL levels, and CRM data, to enhance churn risk prediction accuracy and robustness for financial service organizations. Exploring these modalities and interactions will ensure an effective multimodal fusion learning mechanism to address each modality and synergize their strengths to ensure more precise churn predictions.

Each modality represents a critical component for the broader churn prediction landscape. This thesis will investigate challenges and opportunities posed by each modality to develop holistic understanding culminating in a multimodal fusion learning framework, leveraging the collective power from the three modalities.

Objective 1.1 for modality: speech emotion recognition. Improve paralinguistic emotion detection accuracy and reliability by applying modern speech signal decomposition methods. The objective is to extend current signal based emotion detection methodology boundaries by incorporating novel signal decomposition techniques, ensuring that emotions extracted from vocal interactions accurately reflect member sentiments.

Objective 1.2 for modality: member financial literacy level. Develop a quantitative framework to assess FL levels. This framework will identify key patterns within survey data and large unlabeled financial network datasets with the ultimate goal to build a data-driven and empirical framework to measure financial literacy as a critical predictor of member behavior and churn. This objective underscores the need to enhance member retention strategies.

Objective 1.3 for modality: customer segmentation for churn. Develop a multidimensional framework integrating causal and predictive models, with high-dimensional CRM data, and identify root causes for churn.

1.3 Research Problems

Core problem 1. Considering the primary goal to assess benefits from integrating multiple modalities to predict churn risk, this thesis will attempt to develop a solution based on the following research question.

- To what extent does fusing the diverse modalities, including member FL, paralinguistic emotion detection from vocal interactions, and CRM data, contribute to churn risk prediction accuracy?

The solution to this problem will provide valuable insights into the fusion effectiveness, highlighting potential accuracy gains achievable from predictions based on individual modalities, and also will establish baseline data such as simple demographic data. This foundation enables the integration of complex data modalities, such as financial literacy, emotion detection, and CRM information, to significantly improve churn prediction accuracy.

Core problem 2. Churn risk prediction robustness, which is the model's ability to maintain high levels of prediction accuracy and reliability across various conditions and data variations, depends on successfully integrating individual data modalities into a coherent prediction system. Robustness here aims to integrate different data sources such as financial literacy, emotion detection from voice, and CRM data effectively, ensuring the churn prediction model remains effective and accurate in real-world scenarios where data may be incomplete, noisy, or highly variable. This requires a fusion learning mechanism to consolidate each modality's strengths while mitigating inconsistencies and anomalies. The research challenge is to identify and implement a multimodal fusion learning mechanism that ensures the churn prediction model resilience. The relevant research question is as follows.

- What multimodal fusion learning mechanism offers maximum robustness?

The main research question intricacies are grounded in profoundly understanding each modality. Therefore, each sub-question aims to dissect a specific modality, addressing its challenges and potential, and contribute to answering the main research questions.

Problem 1.1 The first research objective for this thesis focuses on detecting emotions in speech since recognizing the vital role emotions play in member churn and improving their detection accuracy is essential. This will enhance emotion detection by investigating speech signal decomposition contributions to overall churn prediction accuracy. Therefore, this thesis will propose a solution based on the following research question.

- How can para-linguistic emotion detection be improved through speech signal decomposition techniques?

Problem 1.2. Financial literacy is a key aspect for member behavior in financial organizations, but quantitative assessment is complicated. The challenge is to discern meaningful patterns from diverse survey results and unlabeled financial data to measure financial literacy, a significant churn predictor, accurately. The research question guiding this exploration can be expressed as follows.

- How can member FL level be quantitatively assessed by identifying significant patterns in key attributes within survey and unlabeled financial data?

Problem 1.3. CRM data is vast and often high-dimensional. Although this provides valuable insight into historical member data, the curse of dimensionality presents considerable challenges. Therefore, this thesis will attempt to solve the problem based on the following questions.

- How can a work frame be implemented to mitigate impacts from high-dimensional feature spaces in CRM data for more accurate and computationally efficient member churn prediction?

This research question corresponds to the third thesis objective to establish a framework that can efficiently handle sparse CRM data while preserving richness, ensuring accurate and computationally feasible churn predictions.

1.4 Research Core Contributions

The main contribution of this thesis is the development of an innovative multimodal fusion learning framework that significantly advances the fields of churn prediction and member engagement analysis. This research offers a holistic view of member behavior by synergistically integrating diverse data sources, including customer voice interactions, financial literacy assessments, and comprehensive CRM data. This approach enhances the accuracy of churn predictions and provides a more in-depth understanding of the multifaceted nature of member engagement.

- **Advancing speech emotion recognition using hybrid methods.** This thesis will advance speech emotion recognition (SER) by constructing a hybrid acoustic feature map and leveraging previously unexplored Mel spectrogram capabilities. These architectures and methods for data augmentation will constitute substantial progress in emotion recognition technology. *Linked research question 1 and research objective 1.* Improve para-linguistic emotional cue detection by employing advanced speech signal decomposition techniques.

- **Advance financial literacy measurement using SSL in regression.** Develop a more nuanced perspective on FL by introducing mixed methodology SMOGN-COREG to measure FL by regression. This addresses real-valued target variable challenges and will augment current paradigms.

  Linked research question 2 and research objective 2. Develop a framework for holistic understanding through multimodal predictive modeling.

- **Churn propensity modeling framework.** This thesis proposes an innovating churn propensity framework specifically designed for financial behavior analysis, providing a new paradigm for churn evaluation that is adaptable across various business sectors.

  Linked research question 3 and research objective 3. Develop and apply a robust framework for efficiently processing high-dimensional CRM data feature spaces and providing precise churn predictions.

- **Customer churn prediction through multimodal hybrid fusion learning:** This thesis will revolutionize churn propensity analysis by integrating a multimodal fusion approach combining FL, emotion recognition, and CRM data. The proposed method will enhance prediction accuracy and provide a bias mitigation strategy within the customer churn prediction (CCP) model, leading to fairer and more balanced outcomes.

  Linked main research question and research objective. Develop a comprehensive framework utilizing diverse modalities and hybrid fusion techniques for equitable churn prediction in member-centric organizations.

1.5 Ethical Considerations in AI: Safeguarding Privacy in Emotion Detection

The ethical utilization of AI, particularly in detecting emotions from vocal interactions, raises significant privacy concerns. The EU Artificial Intelligence Act, a landmark regulation, addresses these challenges by setting stringent standards for AI applications, emphasizing the protection of fundamental rights and privacy. This Act mandates robust safeguards against the misuse of biometric and personal data, prohibiting untargeted scraping of vocal characteristics for creating recognition databases, and ensuring that any use of emotion recognition technologies adheres to strict privacy and ethical guidelines. Additionally, it restricts AI systems that could manipulate or exploit user vulnerabilities, thereby protecting individuals from privacy infringements and ensuring their vocal data is handled responsibly [11]. In this thesis, we attempt to consider the implications of the EU AI Act, integrating its principles into our research methodology to ensure that our approach to emotion detection not only advances the field but also aligns with these critical ethical and privacy standards. Our approach also aligns with the Australian AI Ethics Principles established in 2019 to uphold democracy, the rule of law, and individual rights, setting a precedent for the ethical management of sensitive data in AI applications.

1.6 Thesis Organization

Figures 1.2 and 1.1 show the thesis organization. This introduction chapter describes the thesis motivation, aims, objectives, and contributions. Subsequent chapters consider the following aspects.

Chapter 2 explores member behavior analysis evolution for churn prediction and real-world applications. Multimodal churn modeling strengths are highlighted by categorizing various learning algorithms. The chapter also summarizes algorithm key features, limitations, and potential avenues for future research.

Chapter 3 considers empirical analysis for factors affecting SER performance, including benefits and limitations regarding pattern classification. SER modality significance for multimodal churn modeling is also investigated along with Mel Spectrogram components to enhance emotion recognition, and acoustic feature augmentation impacts using variational mode decomposition (VMD) based signal decomposition techniques to improve CNN VGG learning. This chapter also compares the proposed VGG-optiVMD algorithm performance against other SER algorithms and provides insight into prospective research avenues for further SER problem exploration.

Chapter 4 introduces the SMOGN-COREG semi-supervised learning framework, designed to label large unlabeled target variable datasets by merging labeled online member surveys with unlabeled transaction datasets. This chapter explains how this proposed methodology overcomes challenges posed by limited sample sizes in online surveys.

Chapter 5 explores causal analysis for member churn within organizations, highlighting its growing importance in the modern complex business landscape. The core aspect is introducing a framework that combines deep feedforward neural network capabilities with insights from sequential pattern mining, specifically regarding high dimensional sparse data inherent to financial domains. The efficacy of this fusion approach compared with existing methodologies, is evident from superior churn prediction outcomes. This chapter also integrates the churn prediction model with a Bayesian causal network, leveraging the DoWhy library, to enhance predictive robustness and provides more comprehensive insights into fundamental churn causes, providing a more comprehensive picture. The chapter concludes by discussing potential directions for future research, focusing on refining churn prediction mechanisms and exploring innovative methodologies to pinpoint underlying causes behind member churn.

Chapter 6 provides a novel perspective on churn prediction within contemporary business environments, addressing limitations with traditional churn models, which depend on a singular data source. This chapter underscores the importance of a multimodal fusion learning approach, advocating for integrating diverse datasets, i.e., CV, FL, and CRM, to develop more accurate churn prediction models.

The SER system is introduced as the first modality in the proposed multimodal modeling to explore member sentiment analysis. Leveraging pre-trained CNN-VGG16 capabilities, this system can proficiently discern member sentiments from vocal attributes such as pitch, energy, and tone. The SMOGN-COREG supervised model adopts FL as a second modality to interpret member FL from historical financial network data. The third modality, i.e., the baseline churn model, is fortified with an ensemble artificial neural network coupled with SMOTE oversampling techniques, to estimate churn probabilities adeptly.

This chapter focuses on innovative fusion techniques, incorporating late and hybrid fusion methodologies into the multimodal method to ensure individual modality feature preservation while optimizing collective synergies to extract holistic insights. Efficacy for this multimodal modeling approach is demonstrated by evaluation metrics and robust prediction accuracy.

Finally, the chapter explains a notable insight from the collated data: a distinct correlation between negative emotion, low FL level, and elevated churn propensities.

Chapter 7 summarizes and concludes the thesis, and discusses practical contributions, key findings, and potential directions for further research.

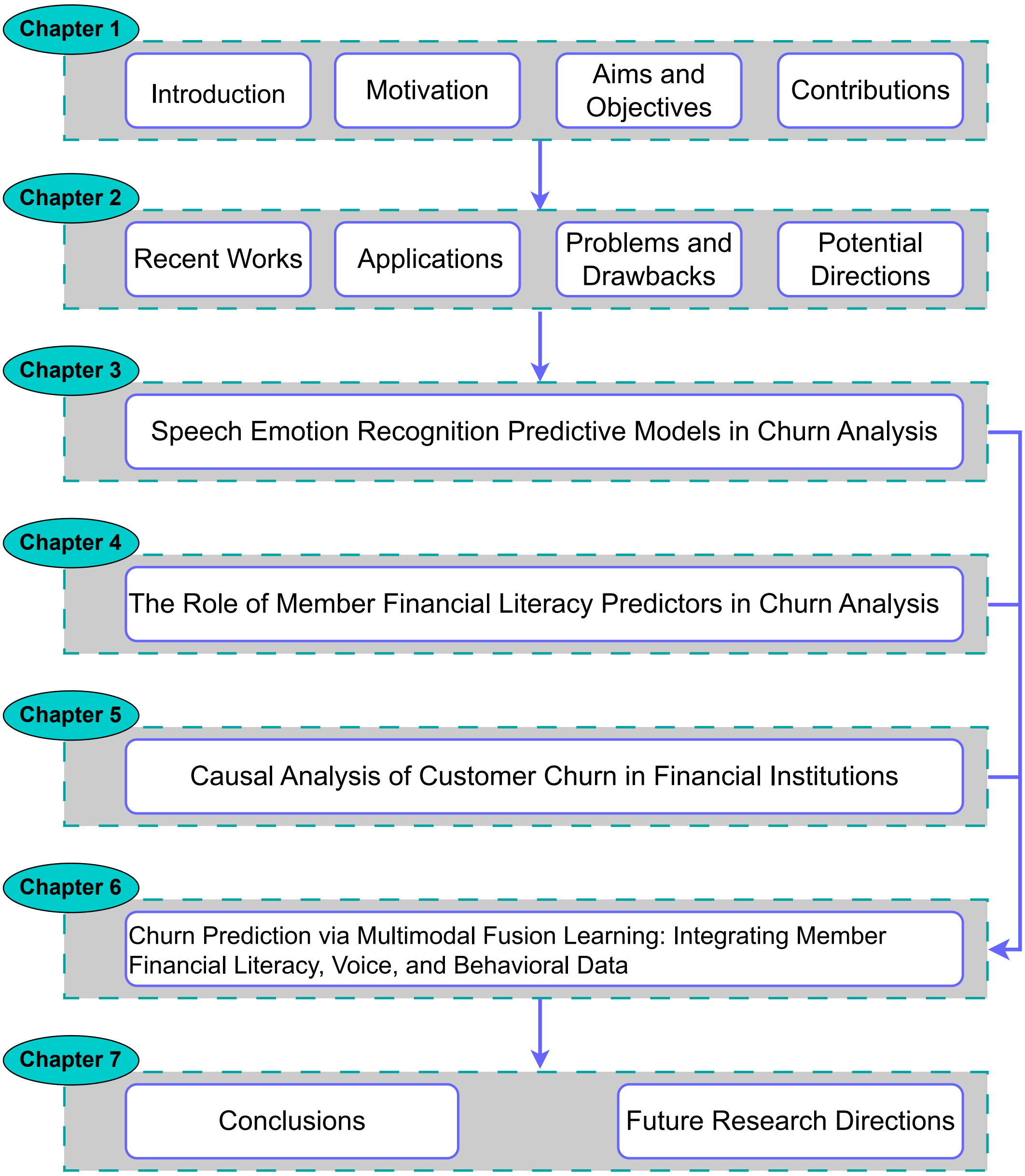

Chapter 1
Introduction
Motivation
Aims and Objectives
Contributions
Chapter 2
Recent Works
Applications
Problems and Drawbacks
Potential Directions
Chapter 3
Speech Emotion Recognition Predictive Models in Churn Analysis
Chapter 4
The Role of Member Financial Literacy Predictors in Churn Analysis
Chapter 5
Causal Analysis of Customer Churn in Financial Institutions
Chapter 6
Churn Prediction via Multimodal Fusion Learning: Integrating Member Financial Literacy, Voice, and Behavioral Data
Chapter 7
Conclusions
Future Research Directions

Figure 1.1: Thesis structure

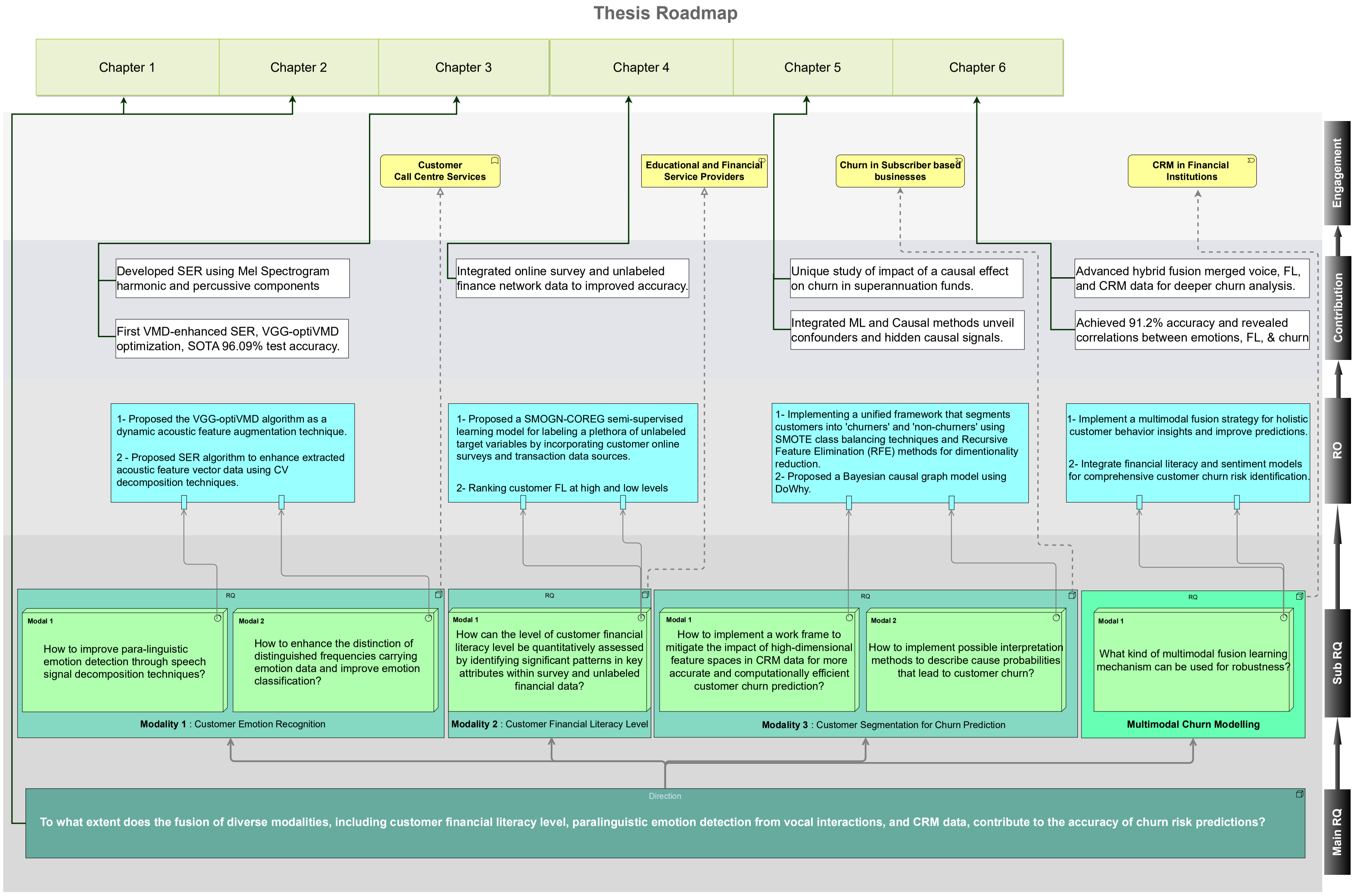

Thesis Roadmap
Chapter 1
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Chapter 6
Customer
Call Centre Services
Educational and Financial
Service Providers
Churn in Subscriber based
businesses
CRM in Financial
Institutions
Developed SER using Mel Spectrogram harmonic and percussive components
First VMD-enhanced SER, VGG-optiVMD optimization, SOTA 96.09% test accuracy.
Integrated online survey and unlabeled finance network data to improved accuracy.
Unique study of impact of a causal effect on churn in superannuation funds.
Integrated ML and Causal methods unveil confounders and hidden causal signals.
Advanced hybrid fusion merged voice, FL, and CRM data for deeper churn analysis.
Achieved 91.2% accuracy and revealed correlations between emotions, FL, & churn
1- Proposed the VGG-optiVMD algorithm as a dynamic acoustic feature augmentation technique.
2 - Proposed SER algorithm to enhance extracted acoustic feature vector data using CV decomposition techniques.
1- Proposed a SMOGN-COREG semi-supervised learning model for labeling a plethora of unlabeled target variables by incorporating customer online surveys and transaction data sources.
2- Ranking customer FL at high and low levels
1- Implementing a unified framework that segments customers into 'churners' and 'non-churners' using SMOTE class balancing techniques and Recursive Feature Elimination (RFE) methods for dimentionality reduction.
2- Proposed a Bayesian causal graph model using DoWhy.
1- Implement a multimodal fusion strategy for holistic customer behavior insights and improve predictions.
2- Integrate financial literacy and sentiment models for comprehensive customer churn risk identification.
RQ
Modal 1
How to improve para-linguistic emotion detection through speech signal decomposition techniques?
Modal 2
How to enhance the distinction of distinguished frequencies carrying emotion data and improve emotion classification?
Modality 1 : Customer Emotion Recognition
RQ
Modal 1
How can the level of customer financial literacy level be quantitatively assessed by identifying significant patterns in key attributes within survey and unlabeled financial data?
Modality 2 : Customer Financial Literacy Level
RQ
Modal 1
How to implement a work frame to mitigate the impact of high-dimensional feature spaces in CRM data for more accurate and computationally efficient customer churn prediction?
Modal 2
How to implement possible interpretation methods to describe cause probabilities that lead to customer churn?
Modality 3 : Customer Segmentation for Churn Prediction
RQ
Modal 1
What kind of multimodal fusion learning mechanism can be used for robustness?
Multimodal Churn Modelling
Direction
To what extent does the fusion of diverse modalities, including customer financial literacy level, paralinguistic emotion detection from vocal interactions, and CRM data, contribute to the accuracy of churn risk predictions?
Engagement
Contribution
RO
Sub RQ
Main RQ

Figure 1.2: Thesis roadmap overview

CHAPTER 2

Related Works

This chapter investigates the relationship between organizational member engagement and their granular financial and emotional behaviors, called financial X-Ray, to identify predictors for churn. This chapter is organized as follows. Section 2.1 describes the comprehensive literature review of relevant previous studies regarding analysis methodologies. Sections 2.2–2.4 discuss member engagement (ME), member emotion recognition (MER), and financial literacy (FL). Section 2.5 investigates causal analysis for member churn (CAMC) and section 2.6 discusses insights from multimodal churn modeling (MCM). Section 2.7 summarizes and concludes the chapter. Each section contributes to a more profound understanding of factors affecting member engagement and attrition.

2.1 Literature Review Methodology

The literature search methodology was designed to ensure comprehensive collection and analysis of relevant scholarly articles. The search spanned multiple databases, including IEEE Xplore, ACM, Science Direct, Google Scholar, arXiv, and Wiley, to cover a wide range of scientific publications. Keywords included "ME in Organizations", "MER", "Speech Emotion Recognition", "Customer Sentiment Analysis", and "Measuring FL Level" to retrieve articles relevant to the research objectives. Inclusion criteria were rigorous, focusing on article relevance to the predetermined topics. The screening process examined the selected article titles, abstracts, keywords, and full texts to filter unrelated studies. This was followed by a critical review process, where the remaining articles were evaluated for their contribution to the field and relevance to the research questions. Screened articles were then classified into categories ME, MER, FL, CAMC, and MCM. The final phase combined data synthesis and content analysis to facilitate extracting significant findings and constructing a coherent narrative around the research topic.

This review critically assessed previous published studies between 2014 and 2021, primarily focusing on premier journals and conferences as classified by the Australian Business Deans Councils (ABDC) and Computing Research and Education Association of Australasia (CORE). Given the absence of a definitive compendium in the field, it prioritized sources rated A/A* by these councils. Select tier B publications were also considered, in particular "Expert System with Applications" and the "International Conference on Customer Behavior Analysis and Computer Human Interface Systems," which are recognized for their impactful citations in data science. The remainder of this section presents the synthesis from these articles to encapsulate trends in member engagement, churn, and associated data mining methodologies.

2.2 Member Engagement in Organizations

Many previous studies have considered member engagement and churn within organizations, proposing several diverse frameworks and analytical techniques [12, 13], with general agreement the FL forms a significant determinant for customer retention. Hastings et al. [14] reviewed relevant financial literacy, financial education, and consumer financial outcome literature, suggesting a strong link between financial knowledge and consumer behaviors relevant to retention. Rudd et al. [8] subsequently proposed a correlation between financial comprehension and member loyalty. Lamba et al. [15] considered telephonic communication, particularly post-pandemic COVID-19, emphasizing call log sentiment analysis predictive power for preempting churn. Some recent studies have applied artificial neural networks to model member behavior in CRM platforms, confirming its efficacy in identifying informative patterns to predict churn [16].

2.3 Member Emotion Recognition Techniques

Assessing member engagement through call logs has become increasingly relevant since the COVID-19 pandemic. Call logs can provide valuable information, including call frequency, duration, and recency. Member emotional tone during calls can be detected by voice analytics, i.e., and subsequently provide a potent indicator for member sentiment and satisfaction and an informative feature for churn modeling.

Previous speech emotion recognition (SER) studies have been significantly influenced by feature extraction and classification technique advances [17]. SER has traditionally been segmented into facial, acoustic, and linguistic domains, with some studies proposing multi-view approaches integrating two or more domains.

Early SER frameworks proposed enhanced support vector machine (SVM) classifiers to predict emotions, such as anger, happiness, and sadness [18–20]. Wu et al. [21] employed traditional machine learning (ML) methods on the EMO-DB database, adding modulation spectral features that amalgamate prosodic features, achieving 85.8% validation accuracy using a multi-class linear discriminant analysis classifier. Milton et al. [22] subsequently integrated three SVMs for emotion classification in EMO-DB, and Huang et al. [23] proposed a hybrid semi-convolutional neural network (CNN) model using deep learning (DL) CNNs (DNNs) to learn feature maps and a traditional SVM to classifying emotions, achieving high test accuracies 88.3%, 85.2%, respectively for both speaker dependent and independent scenarios.

Several recent studies have explored pre-trained CNN image classifier potential, using transfer learning to treat spectrograms as input images, and achieving competitive performance outcomes [24, 25]. Wang et al. [26] introduced the Fourier parameter as an acoustic feature, and Popova et al. [27] achieved 71% accuracy using a fine-tuned DNN and CNN-VGG16 classifier on the RAVDESS dataset. Satt et al. [28] employed a multimodal long short-term memory-convolutional neural network (LSTM-CNN) with a novel feature extraction method based on para-lingual data from spectrograms, achieving 68% accuracy on the IMOCAP database.

Meng et al. [29] proposed a multimodal dilated CNN architecture with a residual block and LSTM by Bai (BiLSTM) to improve classifier accuracy. They achieved remarkable accuracy of 79.96% and 90.78%, respectively, on IEMOCAP and EMO-DB databases. Hajarolasvadi et al. [30] designed a 3D feature framework utilizing a 3D CNN-based classifier; and Zhao et al. [7] proposed a multimodal 2D CNN-LSTM network; both achieved significant accuracy 95.33% and 95.89%, respectively pushing speaker independent classification performance boundaries on the Berlin EMO-DB. Table 2.1 summarizes the various technological approaches for member emotion recognition, showcasing advancements in SER methodologies across different feature extraction methods and learning networks.

Table 2.1: Comparative Overview of Speech Emotion Recognition Technologies.

Model proposed by	Feature extraction method	Learner
Badshah et al. [24]	log Mel spectrogram	CNN
Dendukuri et al. [31]	45d- Mode statistical+MFCCs+Spectral	SVM
Zamil et al. [32]	13 MFCCs	Tree Model
Popova et al. [33]	Mel spectrograms	VGG16
Hajarol. et al. [30]	Mel spectrograms+MFCCs	CNN
Wang et al. [26]	Fourier Parameter+MFCCs	SVM
Kown et al. [34]	Spectrogram	Deep SCNN
Badsha et al. [35]	Spectrogram	CNN
Huang et al. [23]	Spectrogram	CNN
Issa et al. [36]	MFCCs+Chroma.+Mel spec.+Contrast+Tonnetz	VGG16
Meng et al. [29]	log Mel spec.+1st & 2nd delta(log Mel spec.)	CNN-LSTM
Wu et al. [21]	Modulation Spectral Features (MSFs)	SVM
Rudd et al. [37]	Harmonic-Percussive (HP)+log Mel spec.	VGG16-MLP
Demircan et al. [25]	LPC+MFCCs	SVM
Zhao et al. [7]	log Mel spectrogram	CNN-LSTM
VGG-optiVMD	3D-Mel spectrogram+MFCCs+Chromagram	VGG16-VMD

2.3.1 Speech signal decomposition and emotion recognition

Speech signal decomposition has recently experienced significant advancement, with Dendukuri et al. [31] pioneering decomposing speech signals into three distinct components at 16000 Hz over 20 milliseconds. They proposed adding mode central frequency statistical parameters to a SVM classifier, producing several new methods for emotion recognition. Lal et al. [38] subsequently empirically substantiated variational mode decomposition (VMD) efficacy to isolate the correct central frequencies from noise-polluted emotional speech signals, achieving enhanced epoch location estimation. Zhang et al. [39] explored multidimensional feature extraction potential, merging wavelet packet decomposition with VMD for EEG signal emotion recognition. This technique allowed extracting complex features, such as wavelet packet entropy and fractal dimensions, yielding robust classification results when coupled with a random forest (RF) classifier on the DEAP dataset [40]. Khare et al. [41] proposed minimizing reconstruction error through metaheuristic techniques, refining the optimized variational mode decomposition (O-VMD) with 5% increased accuracy on a self-compiled four-emotion dataset.

Pandey et al. [42] proposed combining VMD with DNNs (VMD-DNN) for subject-independent emotion recognition on the DEAP dataset. This enhanced classifier accuracy from VMD based feature extraction that utilized first difference and power spectral density features.

Although previous studies have predominantly utilized STFT signal decomposition techniques for SER, VMD application for speech signal analysis remains relatively new, with most research focusing on EEG signals for emotion recognition research. One objective for the current thesis was to utilize VMD to enhance multidimensional feature vectors and, consequently enhance the VGG16 network [43] learning capabilities in the SER domain. This marks a significant precision improvement in assessing feelings.

As we explore the nuances of member emotion recognition through advanced speech signal decomposition, it becomes evident that emotional intelligence is deeply intertwined with financial behaviors. The emotional states of members, discernible through sophisticated SER techniques like VMD, often reflect their engagement levels and satisfaction with financial services. This emotional feedback is a critical component of broader member profiles, which also encompass FL. Just as emotional dispositions can signal impending churn, a member's financial understanding significantly dictates their financial decisions and long-term loyalty to an institution. Thus, transitioning from the realm of emotion recognition to financial literacy allows us to delve deeper into the psyche of the consumer, where emotional and financial competencies meet to shape overall member engagement and churn.

2.4 Financial Literacy Impact on Member Churn

Enhancing customer FL is essential but significantly challenging for organizations, particularly in the financial sector. Deficient FL among customers often leads to suboptimal product choices and inability to capitalize on financial advisory services. This understanding gap, particularly regarding financial product profitability and utility, can precipitate erroneous decision-making, culminating in reduced organization profitability and increased customer dissatisfaction and churn. Previous FL studies focused on FL surveys (i.e., qualitative methods) or predicting FL (i.e., quantitative methods). The most prevalent method to ascertaining individual financial expertise remains surveys. For example, Worthington [44] conducted an extensive survey encompassing a broad demographic spectrum to correlate FL with socio-economic and demographic traits. Although they employed a logit model to segment FL levels, they encountered significant precision limitations, particularly within median spectrum responses. Previous empirical studies have indicated that lower FL levels are disproportionately prevalent among members resident in socioeconomically disadvantaged areas. Higher education, business ownership, and age all correlate with elevated FL [45]. Huang et al. [46] utilized a back propagation neural network to evaluate FL across diverse financial domains, achieving 92% overall accuracy.

2.4.1 Leveraging unlabeled data in predicting financial literacy

Recent studies have expanded to include semi-supervised learning (SSL), which employs labeled and unlabeled data, in contrast to solely labeled or unlabeled data. Ding et al. [47] advanced this domain with GraphSGAN, an SSL approach using generative adversarial networks on graphs, outperforming conventional methods such as Chebyshev and graph convolutional networks for sensitivity to labeled data [48, 49]. Previous studies have shown that exploiting a small labeled dataset derived from online FL surveys makes it challenging to label the plethora of unlabeled financial network data corresponding to user financial behavior.

2.5 Innovations and Methodologies in Churn Prediction

Churn prediction methodologies have undergone extensive evaluation to identify the most effective techniques [50]. Cutting-edge churn prediction frameworks incorporate DNN models, time-to-event analytics, and big data processing, leveraging GPU computational power for large-scale parallel computing [51]. Employee churn poses similar challenges for organizations, since key customer departures may impose more substantial costs due to the complex nature of the loss compared with employee attrition. However, the repercussions of recruiting and instructing new personnel to replace valuable employees can also generate substantial expenses [52].

Recent advances in predictive modeling have adopted partial least squares (PLS) based techniques, which outperform in generating precise models from highly correlated datasets [53]. The telecom sector in particular has benefitted from hybrid learning algorithms for churn predictive modeling of member behavior [54].

Locally linear model tree methods combine neural networks, fuzzy logic, and decision trees, and the RemsProp training technique has demonstrated superior accuracy compared with conventional algorithms in DL based churn prediction [55, 56]. Randon Forest algorithm efficacy has been well established for in churn prediction, particularly when integrated with sampling methods and cost learning, surpassing many recent well-known algorithms on real banking datasets [57]. Real-world case studies, such as Orange Belgium's customer churn, have been tackled using the ensemble method with an RF classifier to address significant class imbalances [58]. General feature sets extracted from transaction data have been employed in non-subscription business contexts to predict churn using multilayer perceptrons [59], and the CHAMP system was proposed to predict telecommunication service cancellations [60]. Neural networks (e.g. Alyuda Neuro Intelligence) have also been employed for data mining banking customer churn [61].

Adding textual data to CCP algorithms enhances their value [62], and combining classifiers (e.g. gradient boosting) with oversampling techniques has been shown to be effective against skewed data in superannuation funds [63]. Hidden churn factors, which are prevalent in superannuation funds where accounts become dormant, necessitates strategies to improve member engagement and fully utilize member data. Advanced DL techniques have expanded the capacity to manage larger datasets than traditional ML approaches, and integrating DL with CNNs has been successful in predicting churn [64].

2.5.1 Causal inferences for churn

Investigating causal inferences for churn has transformed significantly with recent studies, shifting away from conventional statistical analyses to embrace multivariate causal frameworks [60]. Various innovations, such as Peter Clark (PC) stable algorithm, enable interpreting causal structures from datasets with deep feature sets, facilitating temporal causal modeling for large time series datasets [65]. Directed acyclic graphs are increasingly leveraged within Bayesian networks to depict causal linkages, enhancing predictive accuracy for customer churn in banking [66]. Shah et al. [59] proposed this approach using a model that assigns precise feature weights, which has become instrumental in predicting customer churn across diverse sectors, including telecommunications and finance. Such methodologies have been extensively utilized to discover churn causal variables and construct churn causal models [67]. Lattimore et al. [68] proposed a CNN to gauge sentiment from daily Twitter feeds, verifying the findings with the Granger causality test incorporated in churn models.

Despite these advancements, there remains a gap relating causal analysis of churn, particularly within superannuation funds, associations, and financial institutions. Churn prediction models have been typically examined against datasets from telecommunications, media, and gaming industries. Following on from these previous studies, this thesis proposes a scaled churn prediction methodology. The proposed framework effectively addresses high-dimensional sparse data from local financial institutions with millions of members, combining recursive feature elimination, synthetic minority over-sampling technique (SMOTE), DNNs, and Bayesian causal networks. This mixed-methodology approach addresses the challenge of financial behavioral data collected from CRM platforms for predicting churn.

2.6 Multimodal Churn Modeling

Churn prediction is rapidly moving towards more sophisticated methods, capitalizing on increasing data availability to enhance predictive accuracy. Thus, MCM is also developing rapidly, integrating various customer experience aspects to offer composite understanding for churn dynamics [69]. De Caigny et al. [62] proposed integrating textual data within churn prediction models, highlighting CNN efficacy, outperforming traditional text mining methods to achieve 89.87% accuracy. This advancement verifies that text mining can also assist with predicting churn.

Nhi NY and Liu [70] extended churn prediction methodologies to include unstructured data, such as audio call transcriptions. Their proposed model integrated text mining with CRM data for a gradient boosting tree algorithm, significantly enhancing churn prediction performance across diverse datasets. Kimura [71] demonstrated advantages from blending boosting algorithms with hybrid resampling methods, tackling various challenges due to imbalanced datasets. However, techniques like SMOTE are not yet widely implemented for churn prediction despite their potential. Ahn and Hwang [67] discussed the requirement for adaptable churn prediction methodologies adapted to specific data types, noting a scarcity of diverse data input approaches in existing research.

This thesis addresses the identified gap by employing three distinct datasets to represent multifaceted organizational member engagement for churn prediction. Various hybrid feature fusion approaches are also employed to improve churn prediction in financial institutions by fusing member's emotional feedback from audio calls, historical data from CRMs, and FL survey or financial X-Ray data. Table 2.2 details the proposed methodology, combining diverse data streams, and establishes a new precedent for multimodal churn prediction strategies. Many previous studies have only considered single input data sources, e.g. CRM databases, for churn prediction; whereas integrating diverse data sources is imperative.

Table 2.2: Comparative studies and recent modalities

Ref.	Input	Learning	Prediction	Industry
[72]	1[a]	Unimodal	RF	Finance
[73]	1	Ensemble	GBT+k-medios	Telecom
[58]	1	Unimodal	RF	Telecom
[74]	1	Ensemble	DL+LSTM	Game
[75]	1	Ensemble	LSTM+HS	Game
[70]	2[b]	Feature fusion[d]	GBT	Finance
[62]	2	Feature fusion	CNN+Logit	Finance
Proposed	3[c]	Hybrid Fusion[e]	CNNs+DL	Finance

[a] 1: Structured data, e.g. demographic, account, and CRM data)

[b] 2: Structured + textual data, e.g. call log script and e-messages)

[c] 3: Structured + voice + financial literacy, i.e., qualitative data)

[d] Feature fusion: multimodal feature fusion modeling or early fusion)

[e] Hybrid fusion: multimodal hybrid (early + late) fusion modeling

2.7 Multifaceted Member Engagement Analysis Deficit for Churn Prediction

Previous churn prediction studies have predominantly focused on individual customer data aspects, such as transactional or demographic information, with less attention to the multifaceted nature of member engagement. Despite significant advances in identifying churn predictors within these distinct dimensions, a significant gap remains regarding interactions between member financial behavior, emotional feedback, and overall engagement in predicting churn. Current models have frequently overlooked customer interaction features that cover FL, sentiment from communication channels, and behavioral data. This oversight presents a missed opportunity to understand the full scope for factors influencing a member's decision to remain with or depart from an organization. Current methodologies are advanced within their domains, but fail to consider multifaceted aspects of member engagement that could collectively impact churn.

Thus, a comprehensive churn prediction framework is required to transcend traditional unimodal analyses and incorporate a multimodal approach. Such a model would provide richer, more nuanced understanding of churn by considering how various member engagement forms interact and what that interplay suggests about potential churn. It would also address limitations with current models, which struggle with high-dimensional and sparse modern data sets, particularly for large financial institutions with vast member bases. However, implementing such multimodal strategies in financial applications introduces several technical challenges. Integrating varied data types, such as vocal call interactions, transactional records, and behavioral data, necessitates advanced preprocessing to ensure compatibility across modalities. This integration also involves aligning data with different temporal dynamics and developing a unified feature space that effectively captures crucial inter-modal relationships without information loss. Moreover, the processing and analyzing of extensive multimodal data require robust computational resources to handle the scale and complexity, ensuring the system's scalability and efficiency.

This thesis aims to bridge this gap by adopting a holistic approach considering the full range of member engagement. Therefore, this study constructs a more accurate and predictive churn model integrating data from audio call sentiment analysis, financial transactions, and member surveys. This approach will enhance churn prediction accuracy and provide strategic insights for organizations to better proactively address and mitigate factors contributing to member attrition.

2.8 Summary

This literature systematically explored the multifaceted components contributing to member churn within organizations, mainly focusing on the complex interplay for FL, emotional engagement, and behavioral data. This comprehensive survey highlighted various methodologies and predictive models from traditional statistical analyses to multimodal churn modeling.

Section 2.2 explored member engagement, identifying FL related impacts on customer retention and loyalty. Section 2.3 evaluated member emotion recognition advances and how emotion analytics, particularly from telephonic communications, have become crucial in interpreting member sentiments and forecasting churn. Section 2.4 investigated FL impacts on customer decision making, identifying a critical need for improved FL in preventing churn. Section 2.5 considered causal analysis for member churn, highlighting causal inference method evolution and applications in churn prediction. Section 2.6 fused insights derived from each modality through MCM, confirming the advantages of employing diverse heterogeneous and distinct data sources to achieve holistic understanding of churn.

The review highlighted the necessity for an integrated approach to churn prediction, leveraging the strengths from various data modalities. Section 2.7 identified a current research gap to consider the comprehensive range of member engagement in predicting churn. This thesis aims to address this gap using a proposed hybrid feature fusion technique, providing significant benefits by enhancing predictive accuracy and organizational strategy insights to improve member engagement and reduce churn.

CHAPTER 3

Speech Emotion Recognition Predictive Models in Churn Analysis

This chapter presents a comprehensive empirical study of Speech Emotion Recognition (SER), including

1. detecting positive and negative sentiments and their predictive power for member behavior,
2. impacts from increased negative emotion on member engagement and churn,
3. exploring effects from informative acoustic features in emotion recognition (ER), and
4. comparative analysis for the proposed algorithms against other SER methodologies.

3.1 Background and Motivation

Tone of voice significantly influences conveyed meaning, often more than the words themselves, with facial expressions and vocal variations also playing key roles in expressing emotions [76, 77]. The concept of "mind" extends beyond mere thinking, and includes our emotional states and all unconscious patterns of mental and emotional reactions. Emotions emerge at the intersection of mind and body. They are the body's response to our thoughts and the unconscious mind, effectively mirroring the mind within the body.

Many previous studies have demonstrated that intense emotions can alter the body's biochemical state, hence these biochemical shifts are tangible material manifestations of our emotions. Physical expression of emotions can be observed externally, notably through voice tone and intensity during verbal exchanges. Emotions often signify heightened and energized thought processes and provide essential, albeit subtle, insights into customer satisfaction with the services and products being offered. Therefore, it is crucial to be attentive to customer and member emotions.

Speech emotion recognition is a set of well-known data mining techniques for call center data analytics. The SER domain holds considerable significance across diverse fields, enhancing human-computer interfaces, enriching customer support, and augmenting interactive entertainment and contact center experiences [78]. The intrinsic objective for SER frameworks is to identify distinctive vocal features in varied emotional contexts and hence enrich member engagement with a touch of personalization. CRM teams, for instance, commonly leverage SER to estimate customer or member satisfaction using vocal cues during interactions. Many organizations, from growing startups to tech giants, e.g. Google and Microsoft, are engaged in SER research. Emotional expressions, while universally recognized, are interpreted in ways influenced by cultural norms [79, 80]. In contrast to speech recognition, emotion recognition currently lacks a unified methodology for processing and interpreting emotions from vocal cues [81].

Figure 3.1 shows that member emotion can be categorized as signal or text-based. Text-based systems analyze member sentiments using natural language processing (NLP) techniques, incorporating data from various touchpoints, including social media, call transcripts, emails, surveys, customer reviews, blogs, and forums. Signal-based approaches employ signal-processing techniques to recognize emotional states. Voice-based emotion recognition is employed when we only have access to customer voice (CV). SER is a robust method that cannot be imitated, in contrast with facial expression or text-based sentiment analysis, since these are based on historical call logs. SER systems can also detect member or customer primary and secondary moods by computing negative and positive emotion modes over the call duration.

Various physiological factors can modulate vocal expressions, and sophisticated systems are required to interpret these changes for emotion detection. SER's central challenge lies in extracting distinct and stable features from speech. Such features encompass prosodic elements, e.g. pitch and energy, and acoustic dimensions, e.g. linear predictor coefficients (LPCs), Mel spectrograms, linear frequency cepstral coefficients (LFCC), fast Fourier transform (FFT), chromagram, and Mel-frequency cepstrum coefficients

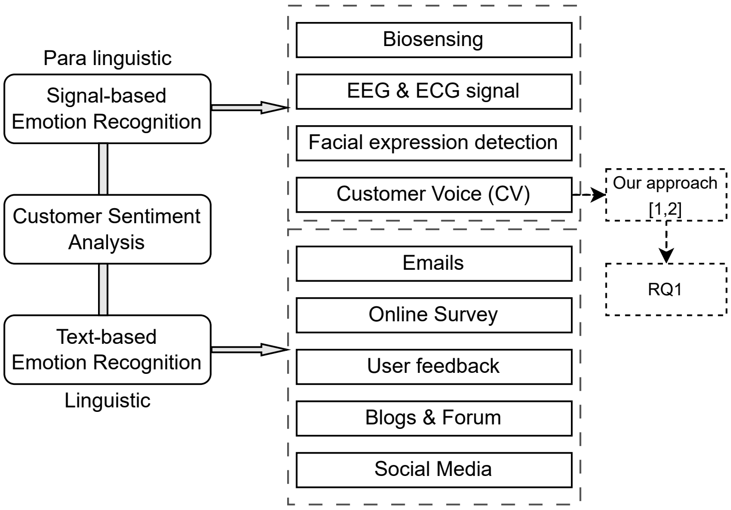


Figure 3.1: Common emotion recognition methods

(MFCCs) [82–86]. Mel spectrogram, MFCCs, and chromagram techniques have been shown to have particular efficacy in extracting emotional data from audio signals [87]. Dual methodologies have also been considered for acoustic feature analysis, some examining proactive features in isolation, and others combining the most informative features to improve model efficacy [88–90]. Empirical findings support data augmentation approaches, fusing prosodic and acoustic feature analysis to ensure a diversified and feature-rich input dataset and hence improving model generalization. Meng et al. [29], Hajarolasvadi et al. [30], and Peng et al. [91] introduced various three-dimensional vocal feature mapping, applying hybrid feature maps to Mel spectrograms and MFCCs for LSTM or CNN-VGG16 [43] extraction.

Following these advancements, this thesis focuses on exploiting the Mel spectrogram for enhanced SER, proposing a novel method that integrates harmonic and percussive elements from Mel spectrograms with the log Mel spectrogram. The core innovation is a hybrid acoustic feature map that improves SER performance, utilizing CNN-VGG16 not only for image analytics but also as a potent tool for emotion classification. Emotions are classified following feature extraction using an optimized MLP network, fine-tuned using a random search hyperparameter sensitivity analysis method, to deliver robust results that parallel current state-of-art accuracy (Zhao et al. [7]).

Speech emotion recognition has substantially evolved from initial reliance on short-time Fourier transform (STFT) methods with RAVDESS, EMO-DB, IEMOCAP, and

WSJCAM databases (amongst others) providing rich resources for training and testing SER models [17, 36, 87]. Integrating CNNs marked a significant SER advance, with pre-trained image classifiers being adapted through transfer learning. Recent advances in this field verify improved efficacy by combining extracted acoustic features from Mel spectrograms and their harmonic and percussive components. However, it is not yet feasible to utilize harmonic and percussive components as a two-dimensional image for CNN input [24, 92, 93].

Although empirical mode decomposition (EMD), wavelet packet decomposition (WPD), STFT methodologies, etc. have been widely employed for electroencephalogram (EEG), electrocardiogram (ECG), and biosensing signal analysis, their application to decomposing customer voice signals for SER remains limited [39, 94]. Applying VMD for speech signal analysis also remains rare, with most studies applying VMD to EEG signals rather than vocal data [31, 38]. Thus, combining VMD with CNNs for data augmentation in acoustic feature extraction remains largely unexplored, but is a potential area for pioneering research [95].

Introducing VMD as a non-recursive signal decomposition method marks a significant advance from EMD and EWT constraints, overcoming their respective limitations and enhancing tone of voice separation performance [8]. Therefore, this thesis leverages VMD in speech signal processing, aiming to extract frequency statistical properties at specific times that distinguish emotions within the feature vector and hence maximize emotion recognition efficacy [31, 41, 96]. The proposed VGG-optiVMD methodology provides a compelling demonstration of how VMD can enrich feature sets, improving classification precision, e.g. recent considerable emotion recognition improvements across well-known emotion datasets.

3.1.1 Key contributions

The main contributions in this chapter can be summarized as follows.

- Proposed an efficient hybrid acoustic feature map technique using harmonic and percussive components from Mel spectrograms, leveraging CNN-VGG16 model strengths, typically used for image processing, to extract and identify emotions from speech signals.

- The first study to employ VMD for dynamic data augmentation in SER, setting a new standard for feature extraction and classification in the field.

- Empirical experiments validate that data augmentation, combining prosodic and acoustic features, significantly improves SER generalization, achieving state-of-the-art 96.09% test accuracy.

3.2 Preliminary Knowledge

3.2.1 Acoustic signal low-level descriptors

Feature extraction in speech analysis typically involves extracting significant features from audio samples and assembling them into extensive vectors. These vectors are subsequently standardized for size using various normalization methods, and often include prosodic and spectral features derived from acoustic low-level descriptors, including the following.

Duration features capture temporal characteristics of speech, including phoneme length, syllables, words, or pauses, and can be normalized in various ways.

Intensity features represent perceived loudness by measuring amplitude over time, mitigating the logarithmic nature of auditory response and spectral distribution's impact on sound perception. They form a loudness contour vector and reflect emotional arousal level.

Pitch features contain data on emotional states due to the tension and vibrations of the vocal cords. Pitch frequency and glottal velocity volume are particularly informative for this purpose.

Formants provide spectral insights into vocal tract characteristics from the frequency and bandwidth of resonances. Emotions can affect sound articulation, causing variations in formant bandwidths. These are typically analyzed using LPCs to estimate formant frequencies.

Spectrum features defined by formants that shape the verbal content. The spectral envelope is assessed using LPCs to compute further characteristics, including centroid, flux, roll-off, and spectral flatness ratio. Long-term average spectra, indicative of overarching spectral patterns, and FFT derived classical spectral elements, offer insights into various parameters, e.g. phase and magnitude. The cepstral domain is also segmented into Mel frequency bands (MFB) to align more closely with human auditory responses.

3.2.2 Acoustic feature extraction

Most previous studies using signal processing techniques for emotion detection focused on extracting statistical features from the signal time-frequency domain, which often holds more information than the time or frequency domains alone. Extracting key spectrum features from speech signal and creating large vectors is a challenging problem for SER. It is important to reshape all the obtained feature vectors to the same size while maintaining the trimmed frame lengths for the most valuable data before utilizing them for model training. Essential features in speech signal processing are Mel spectrograms, chromograms, spectral contrasts, Tonnetz and MFCCs. Mel spectrograms are used in various real-world applications, such as sound event identification [97], speaker recognition [98], and speech recognition [99]. This thesis focuses on leveraging Mel spectrograms using harmonic and percussive components in a hybrid feature engineering technique to improve SER performance. The most important acoustic features in SER can be summarized as follows.

3.2.2.1 Mel-frequency cepstrum coefficients

The speech signal is the convolution of the vocal tract frequency response with a glottal pulse. The most informative data for voice signal processing is vocal tract frequency data, where glottal pulse generated by the vocal cords is considered noise in the speech signal. This thesis used MFCC features to separate these two speech components, employing FFT voice signal mapped onto the Mel scale. The cepstrum result is subsequently obtained from the Mel spectrum by applying the discrete cosine transform (DCT), a simplified FFT, on the log power spectrum. DCT output is coefficient amplitudes, called MFCCs, where their number can be set from 13 to 40 [100].

3.2.2.2 Spectral contrast

Spectral contrast identifies differences between spectral peaks and spectral valleys. Changes in this difference implies a significant change in the emotion behind the voice. Thus, emotional prosody can be decoded, i.e., non-verbal emotional aspects of language [101].

3.2.2.3 Tonnetz

The Tonnetz function is commonly utilized as an alternative representation for pitch and harmony along with the Chromagram and MFCCs. Tonnetz can also estimate tonal centroids on a six-dimensional basis [101].

3.2.2.4 Chromagram

The primary application for chromagrams is to capture harmonic and melodic characteristics of music. However, its use has recently extended to some real-world applications, such as content voice retrieval, song recognition and audio identifiers. For speech signal processing, chromagram features are sensitive to pitch variation in the human voice, providing a powerful tool for emotion recognition. Chromagrams are typically extracted using either the constant-Q transform and or STFT in a defined filterbank. Larger filterbanks provide a high-resolution image of voice data, hence the filterbank setting depends entirely on the application. However, STFT produces a more informative image in a fixed size window; whereas the constant-Q transform provides different data structures in each signal. Thus, STFT chromagrams are easier to synchronize with other features [101].

3.2.2.5 Mel spectrogram

The Mel spectrogram represents the audio signal's frequency spectrum over time, where the frequencies are mapped onto the Mel scale. Thus, a spectrogram is a graphical representation of how the frequency spectrum for a signal changes over time, and is often used to analyze audio. Spectrograms can be displayed in two or three dimensions, with the latter sometimes referred to as a waterfall display. They are crucial for various disciplines, including music, linguistics, and geophysics, for phonetic transcription and animal call analysis tasks. Spectrograms are typically presented as heat maps where colors indicate intensity [102].

Steven and Volksmann [103] established that humans hear different sound frequencies in a nonlinear manner, i.e., the human ear can detect the distance between lower frequencies better than higher frequencies, hence increasingly large intervals are judged above 500Hz by voluntary listeners. The authors proposed the Mel unit, which mitigates nonlinear detection to provide pitch sounds equally distant to the listener, which can be expressed as

$$f_{mel} = 2595.log(1 + \frac{f}{700Hz}), \tag{3.1}$$

where f denotes the input audio signal and f_{mel} represent the converted f to Mel band.

The Mel spectrogram was subsequently synthesized as follows

$$LMS(m) = \sum_{k=f(m-1)}^{f(m+1)} log(H_m(k) \ . \ |X(k)|^2), \tag{3.2}$$

where $|X(k)|^2$ is the power spectrum within the kth frequency bin, k is the index associated with the FFT, m is the MFB quantity, and LMS is the logarithmic Mel spectrogram.

3.3 Methodology

This section discusses two distinct approaches to enhance SER performance. The initial method investigates Mel spectrogram acoustic features', along with their harmonic and percussive components, effects on SER efficacy. The subsequent method explores decomposition based speech signal processing, which helps to identify emotional states from member or customer vocal expressions.

3.3.1 Method 1: Speech emotion recognition using Mel spectrogram harmonic and percussive components

This approach computes the average harmonic and percussive components for the Mel spectrogram and combine the result with the log Mel spectrogram. The proposed framework's efficiency was compared with previous studies and other comparable models that employed different data augmentation methods.

Voice samples were extracted from recorded voice files before implementing feature extraction, ranging from four seconds duration at 88 kHz sample rate. To guarantee frequency resolution and minimize spectral leakage, these samples were then digitized and processed using the Hanning window function [104],

$$H_m(k) = 0.5[1 - cos(\frac{2\pi.k}{M-1})] = sin^2(\frac{\pi.k}{M-1}) \qquad 0 =< \ k \ < \ M-1, \tag{3.3}$$

where M denotes the number of sample points in the output window, k present specific FFT used in Hanning window function. The method utilizes the Librosa library [105] for feature extraction and accordingly sets Mel filter banks, window, and hop lengths.

The first feature map was obtained by applying the log Mel spectrogram (3) to measure the Mel spectrogram output sensitivity to changes in voice signal amplitude. Figure 3.2 shows a representative log Mel spectrogram depicting diverse emotions from the EMO-DB dataset, exhibiting distinct amplitude and frequency representations for each emotion.

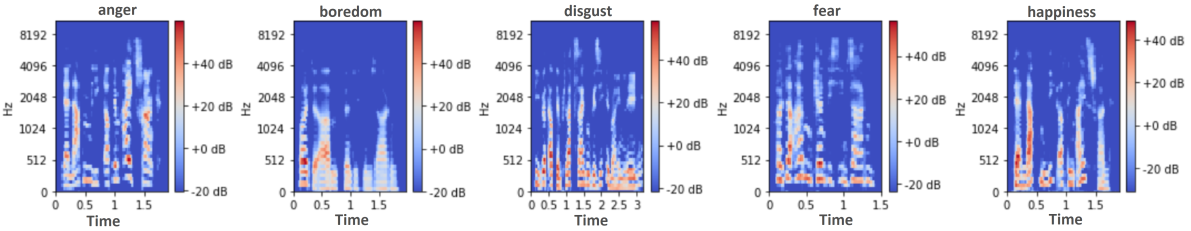


Figure 3.2: Mel spectrograms of voice signal clearly illustrate amplitude and frequency difference for each emotion. Frequencies that contribute more than orange and white colours are shown in red.

Decomposition is pivotal role for extracting harmonic and percussive components from an audio signal, which is critical for enhancing SER performance. This process begins with applying the STFT to the audio frames, yielding spectrogram S from the input signal s,

$$s \; = \; S_h \; + \; S_p, \tag{3.4}$$

and

$$S(n,\, k) \; := \; \sum_{r=0}^{N-1} s(r+nH) \; . \; \omega(r) \; . \; e^{(\frac{-2\pi.kn}{N})}, \tag{3.5}$$

where S is the spectrum obtained from input signal s, ω is the sine window function which defines the window length, H is hop size, n is the current frame number, and N is the FFT length applied to each frame.

Applying median filtering along the time (horizontal) and frequency (vertical) axes for S separates the harmonic S_h and percussive S_p component,

$$\widehat{H} = \widehat{S} \;\bigotimes\; M_H, \tag{3.6}$$

and

$$\widehat{P} = \widehat{S} \;\bigotimes\; M_P, \tag{3.7}$$

obtained from

$$\mathrm{F}_{2(LMS)} \; = \; \frac{(\widehat{H} \; + \; \widehat{P})}{2}, \tag{3.8}$$

where $F_{2(LMS)}$ represent second feature map resulted by mean of harmonic $\widehat{H}$ and percussive $\widehat{P}$ components.

The harmonic component S_h captures tonal elements and the percussive component S_p captures the rhythm and instantaneous nature of the audio signal.

This decomposition technique allows the SER system to distinguish between different emotional expressions that might be conveyed more strongly in either the tonal or rhythmic aspects of speech. Actual implementation requires a more complex set of operations, including specific definitions for the median filters and how they are applied to the spectrogram. Fitzgerald [93] indicated a specific method for spectral decomposition.

Figure 3.3 shows the harmonic and percussive component feature map, built using the average of these components from the Mel spectrogram. Components are separated using a spectral decomposition process adapted from Fitzgerald [93], where the harmonic and percussive elements were distinguished by applying median filters in time and frequency directions on the spectrum.

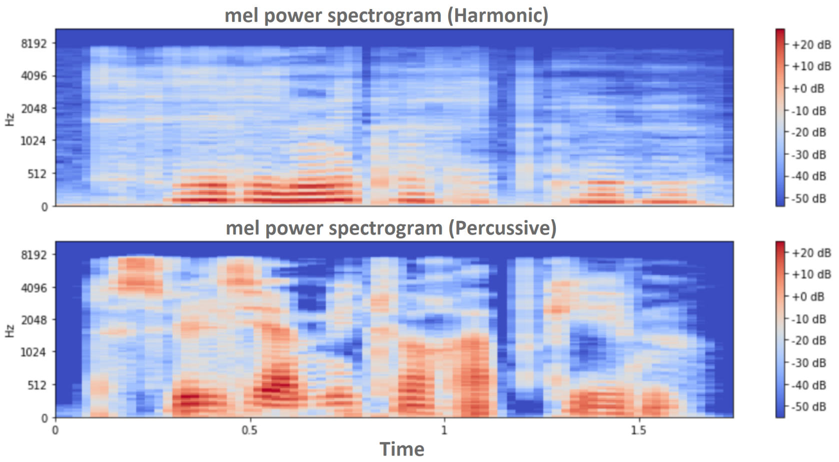


Figure 3.3: Harmonic and percussive components for Mel spectrograms for neutral emotion

Figure 3.4 shows the hybrid feature map, i.e., average the two extracted features, forming a (128 * 128 * 2) combined 2D feature as input data for VGG16. The proposed hybrid feature map function represents acoustic features necessary for training CNN-VGG16 networks. Based on empirical experiments, this particular feature combination demonstrates its efficacy in emotion prediction.

The recent concept to use pre-trained networks in SER considerably improved computational processing of emotion in speech [93]. Therefore, we can build a fresh CNN network to process voice features and then feed the results to an MLP classifier. Using a pre-trained CNN, such as VGG-16, as a feature extractor is an effective approach for SER feature analysis, exploiting CNN-VGG16, which was already trained on the ImageNet

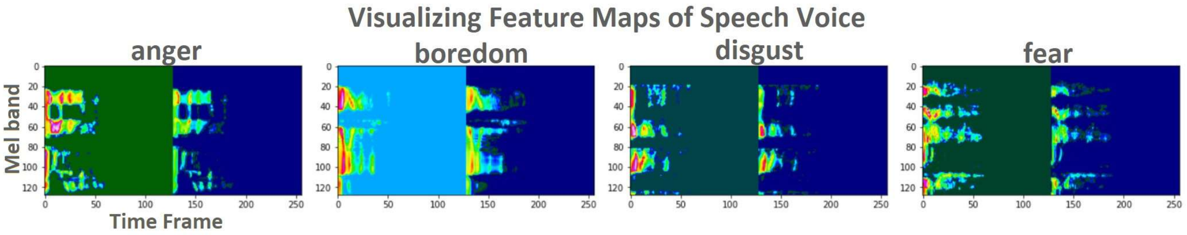


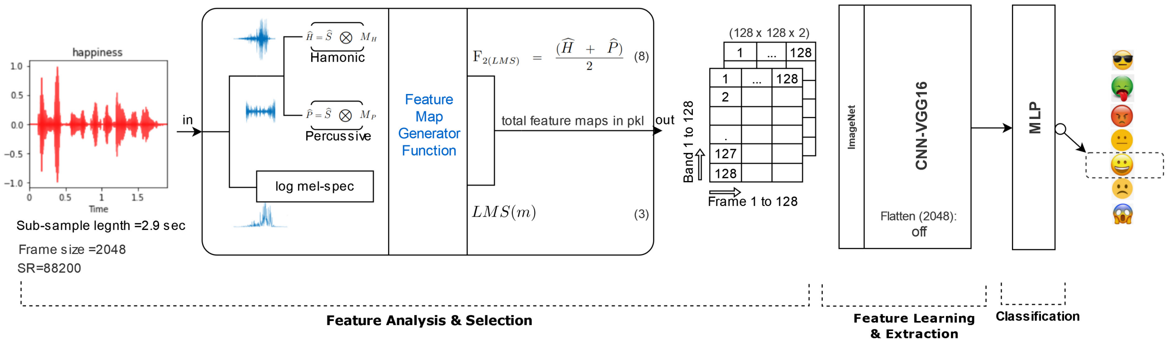
Figure 3.4: Hybrid feature map output visualized in 2D.

dataset [43]. The experimental results confirm that transfer learning models can be generalized in voice signal processing. The example used CNN-VGG16 as a feature extractor only and deactivated its dense layers since it was not used as a classifier.

The drawback of transfer learning networks is the complexity and difficult result interpretation. However, analyzing the activation functions in image processing neural networks improves understanding of how vision-based deep learning models recognize object shape and edge in an input image [106]; this ability can be applied to a spectrogram image. The CNN-VGG network's ability to recognize fine details in high-dimensional feature maps was essential to make this application work for subtle differences between features. Although the VGG16 network's requirement for substantial memory storage poses a limitation for straightforward classification tasks, its utility in comprehensive feature analysis is undeniable.

Figure 3.5 shows the proposed architecture, and integrates the VGG16 network with an MLP network functioning as a feature extractor and emotion classifier, respectively. Feature maps were built from subsamples within a predefined window size, forming a 2D image feature reshaped for size in 128 frames and bands. VGG16 receives these $(128 * 128 * 2)$ feature maps as input image data, outputting a 2048-dimensional vector to feed into the MLP classifier. This classifier was structured with four fully connected layers, employing the ReLU activation function with a softmax output layer.

The proposed framework enhances SER by separating the Mel Spectrogram into harmonic and percussive elements, capturing emotional cues in voice signals. Improved SER accuracy facilitates more profound insights into member sentiments during interactions, a crucial indicator for engagement levels. By accurately gauging emotional responses, organizations can better predict and address causes of member churn, thereby making more targeted retention strategies and preventing member churn.


happiness
1.0
0.5
0.0
−0.5
−1.0
0
0.5
1
1.5
Time
Sub-sample legnth =2.9 sec
Frame size =2048
SR=88200
in
$\hat{H} = \hat{S} \otimes M_H$
Hamonic
$\hat{P} = \hat{S} \otimes M_P$
Percussive
log mel-spec
Feature Map Generator Function
$F_{2(LMS)} = \frac{(\hat{H} + \hat{P})}{2}$ (8)
total feature maps in pkl
$LMS(m)$ (3)
out
(128 x 128 x 2)
1
...
128
2
127
Band 1 to 128
Frame 1 to 128
ImageNet
CNN-VGG16
Flatten (2048): off
MLP
Feature Analysis & Selection
Feature Learning & Extraction
Classification

First Feature:

$$S(n,\ k) \ := \sum_{r=0}^{N-1} s(r+nH) \ . \ \omega(r) \ . \ e^{(\frac{-2\pi.kn}{N})} : \begin{cases} \hat{H} = \hat{S} \otimes M_H & \text{horizontal median filtering} \\ \hat{P} = \hat{S} \otimes M_P & \text{vertical median filtering} \end{cases}$$

Second Feature:

$$LMS(m) = \sum_{k=f(m-1)}^{f(m+1)} log(H_m(k) \ . \ |X(k)|^2)$$

Figure 3.5: Method-1 workframe: leveraging Mel Spectrogram by harmonic and percussive components to improve SER performance.

3.3.2 Method 2: VGG-optiVMD extended VMD algorithm for SER

3.3.2.1 Speech signal decomposition with VMD

Variational mode decomposition is a prominent method to decompose nonstationary signals into discrete sub-signals or modes. Each mode captures distinct characteristics from the original signal within a limited bandwidth centered around a specific frequency. These modes were extrapolated from Hilbert transform outputs, called intrinsic mode functions (IMFs), accurately reflecting the signal's actual components for sufficiently narrow bandwidths [38]. Adaptability of the VMD algorithm is crucial for simplifying the original signal's complexity, facilitating a more focused analysis [95, 107].

The VMD process integrates Wiener filtering, Hilbert transformations, analytical signals, and frequency mixing techniques. The Wiener filter is primarily a narrowband filter for denoising [95]. The Hilbert transformation is a linear time-invariant operator that convolves the original signal $g(t)$ with $1/\pi t$, converting the real signal into a complex signal that helps extract the magnitude and phase time series for influential frequencies at particular time instances [108]. Although the Hilbert transform is theoretically only applicable to narrowband nonstationary signals, it exhibits remarkable performance when combined with a finite impulse response bandpass filter. The VMD algorithm enhances the original signal $g(t)$ by incorporating its Hilbert transform $\mathscr{H}[g(t)]$, effectively eliminating any negative frequency band due to Hermitian symmetry. As demonstrated in Equation (3.9), the frequencies are subsequently mixed by multiplying

$\omega_1 \times \omega_2$, resulting in a nuanced examination of the signal's intricate properties,

$$2\cos(\omega_1 t)\cos(\omega_2 t) = \cos((\omega_1 + \omega_2)t) + \cos((\omega_1 - \omega_2)t) \tag{3.9}$$

where ω_1 and ω_2 indicated target frequencies should bypass via Wiener filter. The analytical signal formulation employs the Hilbert transform, $\frac{j}{\pi t}$, to convert the real-time voice signal into the complex domain. This transformation is essential to extract the instantaneous signal frequency and amplitude. The Hilbert transform is coupled with the unit impulse $\delta(t)$ or Dirac delta function. This function acts as an "impulse" that is infinitely high at the signal's origin point and zero elsewhere. It is a mathematical concept used in signal processing to isolate a single point in time. Applying the Hilbert transform to a signal essentially shifts the phase of all frequency components by 90 degrees. This phase shift is pivotal for constructing the analytic signal, combining the original function with its Hilbert transform to create a complex signal whose real part is the original signal and imaginary part represents the phase-shifted version. The challenge of reconstructing the original voice signal from its transformed state involves solving a constrained optimization problem,

$$\begin{gathered}\min_{\{u_k\},\{\omega_k\}} \left\{ \textstyle\sum_{k=1}^{K} \left\| \frac{\partial}{\partial_t} \left[\left(\delta(t) + \frac{j}{\pi t} \right) * g_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \\ \text{subject to: } g(t) = \quad \textstyle\sum_{k=1}^{K} g_k(t), \end{gathered} \tag{3.10}$$

where operation $\frac{\partial}{\partial_t}[.]$ is employed to reduce bandwidth variations for the signal modes extracted; $g(t)$ is the discrete frame for the original speech signal being analyzed; each $g_k(t)$ signifies the kth mode derived from $g(t)$, providing a granular view of the signal's characteristics; K is the total number of modes, which encompasses the complete set of extracted signal components, where $\omega_k = \{w1, \ldots, wk\}$ is the central frequencies of these modes.

These central frequencies provide a pivotal reference, capturing the dominant frequency components for each mode. Additionally, $e^{-j\omega_k t}$ acts as a modulator function, which is instrumental in translating the frequency spectrum for each individual mode down to the baseband. This modulator function allows the bandwidth located around the central frequency ω_k.

This particular type of issue aims to identify the optimal resolution within predetermined parameters or constraints. A Lagrangian multiplier is introduced to streamline this procedure and enhance its computational feasibility. This transforms the problem into an unconstrained optimization,

$$\begin{aligned}\mathscr{L}(g_k,\omega_k,\lambda) := \alpha \textstyle\sum_{k=1}^{K} \left\| \frac{\partial}{\partial_t} \left[\left(\left(\delta(t) + \frac{j}{\pi t} \right) * g_k(t) \right) e^{-j\omega_k t} \right] \right\|^2 \\ + \left\| g(t) - \textstyle\sum_{k=1}^{K} g_k(t) \right\|_2^2 + \left\langle \lambda(t), g(t) - \textstyle\sum_{k=1}^{K} g_k(t) \right\rangle,\end{aligned} \tag{3.11}$$

where, λ is a time-dependent Lagrangian multiplier, and α is a bandwidth control parameter. This effectively removes the constraints by incorporating them into the optimization objective, allowing for a more straightforward solution that optimizes signal reconstruction while satisfying the original constraints.

To extract intrinsic mode functions (IMFs) and their corresponding central frequencies from the speech signal, the unconstrained Lagrangian problem (3.11) is tackled using the alternate direction method of multipliers [95, 109, 110], to facilitate decomposition in the spectral domain. The equivalent optimization outcomes are the same when applied in either the frequency or time domain. Consequently, mode $g_k(\omega)$ associated with each IMF can be iteratively upgraded within the spectral domain,

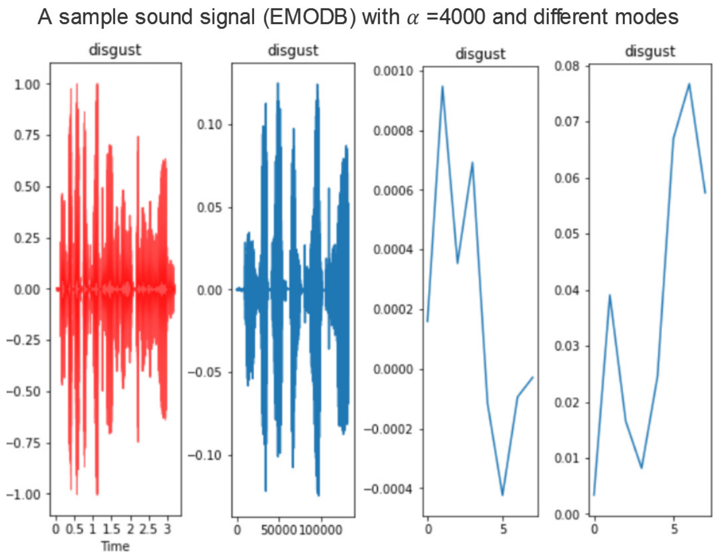
$$\hat{g}_k^{n+1}(\omega) \leftarrow \frac{\hat{g}(\omega) - \sum_{i<k} \hat{g}_i^{n+1}(\omega) - \sum_{i>k} \hat{g}_i^{n}(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha \left(\omega - \omega_k^n\right)^2} \tag{3.12}$$

where $g_k(\omega)$ is frequency domain of decomposed k_th mode of input signal.

The Wiener filter is applied to the residual signal, utilizing the signal $1/(\omega - \omega_k)^2$, which limits oscillation around the central frequency minimum. This constraint assists in stabilizing the frequency spectrum around each mode's center, providing an updated and more accurate estimation for the central frequency ω_k for each mode as indicated in (3.13). This update is a crucial step for variational mode decomposition, since it ensures that each mode is finely tuned to its specific frequency band, capturing the essential characteristics from the original signal necessary for accurate decomposition,

$$\hat{\omega}_k^{n+1} = \frac{\int_0^{\infty} \omega \left| \hat{G}_k(\omega) \right|^2 \, \mathrm{d}\omega}{\int_0^{\infty} \left| \hat{G}_k(\omega) \right|^2 \, \mathrm{d}\omega}, \tag{3.13}$$

where $\hat{G}_k(\omega)$ is the FFT for the newly updated mode $g_k^{n+1}(t)$ at iteration $n+1$. This transformation shifts the updated mode from the time domain into the frequency domain, enabling analyzing and processing the signal based on its frequency content. Accurate transformation of these modes is essential for capturing speech signal characteristics indicative of the emotional states, which is critical for SER systems. Figure 3.6 shows signal decomposition in three modes.



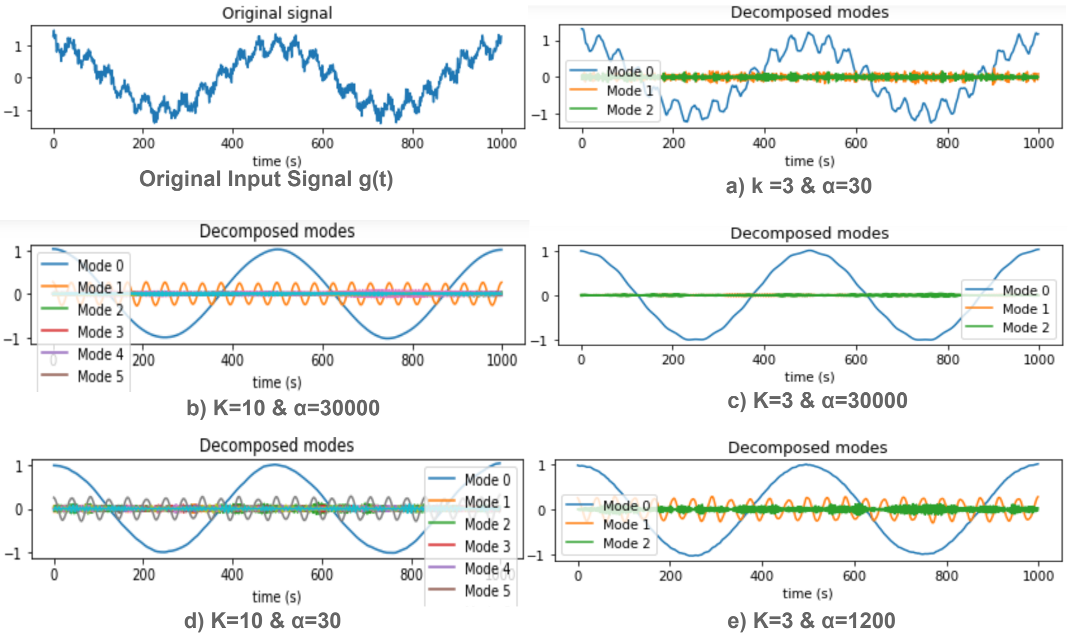
(a) Original (b) Mode-1 (c) Mode-2 (d) Mode-3

Figure 3.6: Speech signal decomposition in three modes.

3.3.2.2 Challenges and advances in speech signal decomposition:

Decomposition for a nonstationary input signal among multiple sub-signals is not unique since the mode has AM-FM signal format due to the two individual harmonics [95]. Reconstruction error for a decomposed signal can be reduced by selecting optimum K and α. However, finding the optimal values empirically is time-consuming and unreliable because incorrect K and α selections can cause information loss from sub-signals, negatively affecting the learning process and reducing classifier performance. Figure 3.7 shows an example where decomposition outcome is more dependent on the band constraint, controlled by α, hence large K and small α can duplicate the noisy sub-signal. However, too small K and too large α causes mode mixing eliminating information within the sub-signals, and the output becomes a micro and macro segmentation of the signal data that no longer contains meaningful features.

The one drawback of VMD is the difficulty to determine optimal decomposition parameters K and α. Several approaches have been proposed for ER using ECG, EEG, and vibrational signals. For example, the optimal VMD algorithm (O-VMD) [111] uses a series of indicators, including permutation entropy, kurtosis criteria, extreme frequency domain value, and energy loss coefficients, to identify optimum K. Wang et al. [112] controlled power spectral and dynamic entropy features to find optimal K and α to decompose vibration signal and extract fault features.



Original Input Signal g(t)

a) k =3 & α=30

b) K=10 & α=30000

c) K=3 & α=30000

d) K=10 & α=30

e) K=3 & α=1200

Figure 3.7: Typical speech signal decomposed over different modes. Various K and α parameter sets capture different nonstationary signal properties: a) too small K and α causes under-segmentation of noisy sub-signals with mode overlap; b) too large K and α captures macro-segmented data from the input signal and distribute informative signal data to different modes; c) too small K and too large α causes neighboring mode interference, with important signal data distributed to different modes; d) too large K and too small α causes over-binning and duplicate signal modes with improper decomposition structure; e) optimum $K = 3$ and $\alpha = 1200$.

However, these approaches use IMF or mode characteristics to find the best decomposition parameters for specific low amplitude input signals with empirical threshold selection, which is not applicable for speech signal processing. Dendukuri et al. [31] decomposed speech signals using five modes to recognize eight emotions, achieving 61.2% accuracy on the RAVDESS database. They combined different features, constructing a 45-dimensional feature set including mode center frequency, statistical values for mode center frequency, MFCCs, and spectral statistical features to improve classifier performance.

The above methods evaluate optimum K using statistical features and indicators for guidance. In particular, identified mode number correctness was not verified or fine-tuned practically by monitoring classification accuracy. Improper decomposition parameter selection will create duplicate modes, causing signal information losses and hence reduced classifier performance.

In contrast, this thesis proposes a framework to automate optimum VMD decomposition parameter selection using a feedback loop from the VGG16 flattening output layer. The optimized VMD algorithm (VGG-optiVMD) is specifical for audio signal processing. The key strengths for VGG-optiVMD are reliability, generality, and reproducibility across different speech databases for real-world applications.

3.3.2.3 VGG-optiVMD: advances in signal decomposition techniques

This thesis proposes the VGG-optiVMD algorithm dynamic acoustic feature data augmentation by extending the current variational mode decomposition algorithm. VGG-optiVMD can enhance frequencies distinction carrying paralinguistic emotion data and improve SER performance. Input data for VGG-optiVMD combines the Mel spectrogram, MFCCs, and chromogram data frames extracted from a speech signal. This acoustic data frame is decomposed into dynamic modes using the $K(2-6)$ and $\alpha(2000-6000)$ parameters. Each mode data frame is concatenated after decomposition and embedded into a larger data frame, creating an augmented acoustic data frame. This approach enhances informative emotion data in acoustic feature maps, and can consequently boost VGG16 training by providing these augmenting acoustic data frames.

This specific approach is the first to utilize VMD as a dynamic acoustic feature data augmentation for SER. Another outstanding VGG-optiVMD ability is to automatically select the VMD decomposition parameters K and α, which guarantees the most optimum emotion classifier performance. This is achieved due to iterative tuning parameters K and α by the VGG-optiVMD algorithm until maximum accuracy (ACC) and F1 score are achieved in the VGG16 classifier. The algorithm sets initial K and α, then changes them iteratively while observing classification accuracy until it obtains the highest AUC and F1-Score metrics, or reaches the break loop condition. The algorithm automatically and effectively selects the optimal decomposition parameters based on a diverse set of K and α testing, ultimately converging on the highest model performance, rather than relying on decomposition parameter convergence.

Figure 3.8 shows that model development commences with sampling the voice signal at 88,400 Hz, and deriving five prominent acoustic features within the time-frequency domain, i.e., MFCCs, Mel spectrogram, Tonnetz, spectral contrast, and chromagram. The Hann window function is then applied to the sub-signal spectra, with fixed length = 2.9 s and shifting time = 0.4 ms across a sequence of frames. Extracted features are then consolidated into a unified feature vector with dimension ($128 \times 128 \times 3$). The SMOTE [113] oversampling technique is applied to enhance minority classes representations and mitigate model bias. The VGG-optiVMD algorithm is then applied to extract frequency statistical characteristics at precise temporal instances, crucial for differentiating emotions within the feature vector. The culmination is training the VGG network on the augmented feature vector, enabling emotions to be classified into seven distinct categories. The extracted features undergo enrichment via VGG-optiVMD, which intuitively identifies optimal K and α, ensuring a refined and accurate emotion recognition performance.

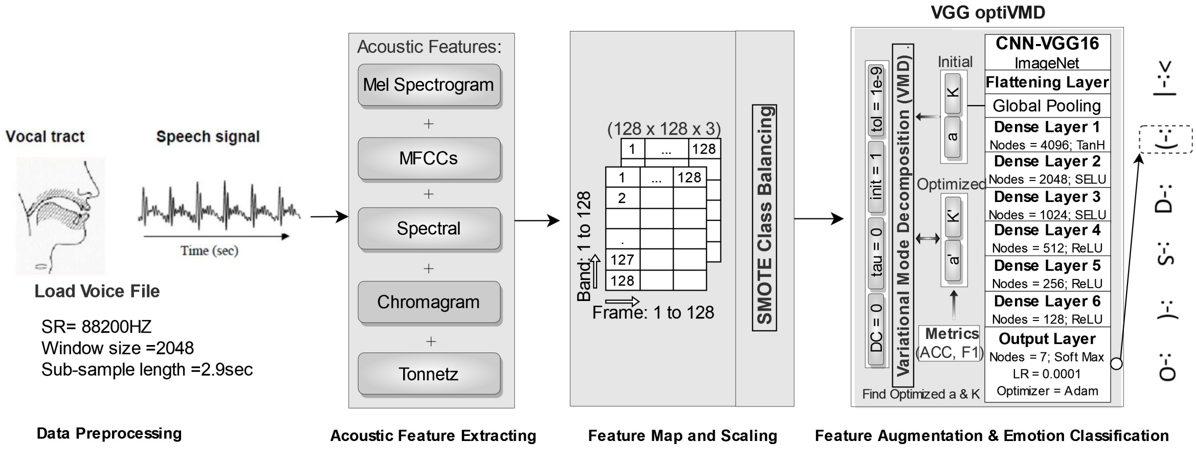


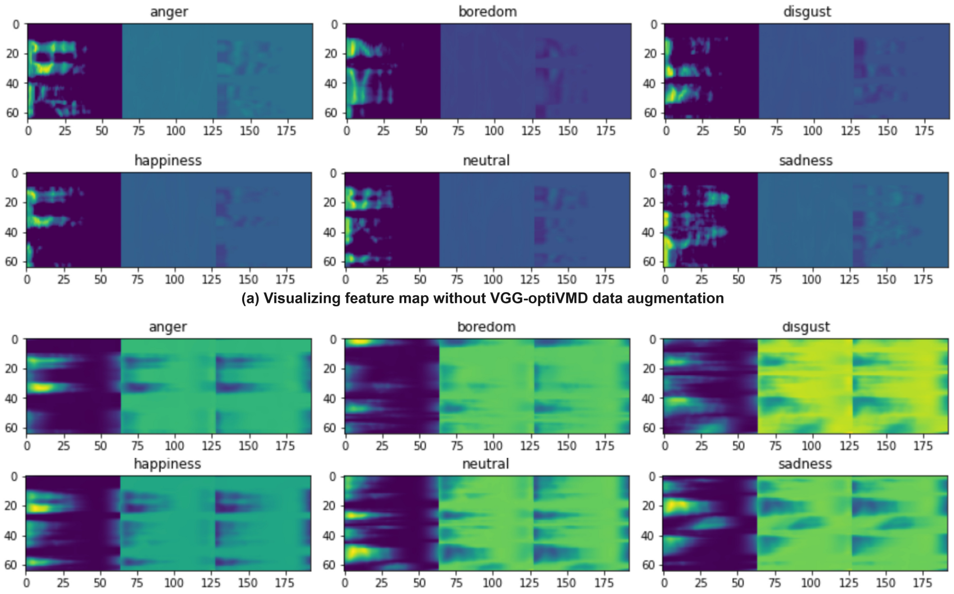
Figure 3.8: Optimizing emotion recognition classification from signal data augmentation to VGG16 network training using VGG-optiVMD.

Figure 3.9 shows the efficient functionality for VGG-optiVMD on the feature vector 3D-Mel spectrogram+MFCCs+chromagram. Figure 3.9(a) and (b) show the initial feature state prior to augmentation; and the enriched frequency distinguished on the augmented feature map data frame achieved post-augmentation, with a markedly increased energy distinction, respectively.

3.4 Experiment Outcomes

3.4.1 Material

Customer voices (CVs) were mapped with similar emotion sample voices from standard emotion databases to maintain privacy while preserving emotional content in the recorded voice files from inbound calls. This de-identifying technique constructed a shadow CV database utilizing the correlation between negative emotions and high-risk churn customers with low FL, and the association between positive emotions and low-risk



(a) Visualizing feature map without VGG-optiVMD data augmentation

(b) Visualizing feature map with VGG-optiVMD data augmentation

Figure 3.9: Enhancing spectral feature discrimination: VGG-optiVMD's proficiency with 3D Mel spectrogram, MFCCs, and chromagram.

churn customers. The Berlin EMO-DB database, a standardized resource for categorizing emotions based on voice recordings, was utilized to label customer motions [114].

3.4.1.1 The Berlin EMO-DB dataset

The Berlin EMO-DB, is a well-known database frequently used for SER, and contains 535 audio files in WAV format. These audio files are categorized into seven emotions: neutral, fear, anger, happiness, sadness, disgust, and boredom. The Berlin EMO-DB was constructed by five female and male actors between 25 and 35 years old who were asked to read ten prepared texts while performing in seven different emotions [114].

Table 3.1: Sample voice distribution in the Berlin EMO-DB dataset

neutral	anger	fear	happiness	sadness	disgust	boredom
79	128	68	71	62	46	81

3.4.1.2 RAVDESS database

The Ryerson audio-visual database of emotional speech and song (RAVDESS database) was employed for model comparison. RAVDESS is a validated, balanced emotional speech and song collection by 24 actors encoded in a neutral North American accent with lexical consistency. It provides a range of emotional states across two intensity levels, plus a neutral baseline, in speech (calm, happy, sad, angry, fearful, surprise, disgust) and song (calm, happy, sad, angry, fearful). Available in face-and-voice, face-only, and voice-only formats, the database includes 7356 multiply rated recordings for emotional validity, intensity, and genuineness by 247 North American raters. The dataset exhibited high emotional validity and interrater reliability, corroborated by test-retest data from 72 participants. Enhanced with corrected accuracy and metrics, RAVDESS facilitates precise stimuli selection for emotional research [113].

3.4.2 Method 1: Harmonic-Percussive Mel spectrogram

The voice samples were divided randomly, with 80% allocated for training and the remaining 20% split evenly between validation and testing. Voice sample imbalance across the seven emotion classes was addressed by employing an oversampling strategy to increase minority class voice samples. Window size was set to 2048 and configuring (128 × 128) bands and frames produced 167,426 signal subsamples and 9,717 feature maps at 88 kHz sample rate. The Librosa toolkit was employed [105] to extract Mel spectrogram features with Mel filterbanks size, window, and hop length = 128, 2048, and 512, respectively.

Foundational feature representations were then constructed using the hybrid feature map extractor function and transformed into 2D image input data. Training data was applied to the VGG16 network to recognize the intricately designed hybrid feature maps. An MLP classifier was subsequently employed to predict seven emotions from the 2048-long one-dimensional vector created by VGG16. The random search method was employed to estimate sensitivity for the optimum MLP model configuration, and the output layer MLP included four fully connected layers, incorporating ReLU activation and softmax functions. The first two dense layers were configured with 1024 inputs and dropout =0.5, whereas the latter two layers were configured with 512 inputs and dropout = 0.3. The ADAM optimizer was selected to optimize the MLP network with learning rate = 0.0001. The classifier was trained over 128 epochs with batch size = four, taking advantage of the NVIDIA GPU computational power.

I further evaluated The proposed hybrid feature method was compare with several traditional acoustic feature extraction techniques, with different MFCCs, chromagram, Tonnetz, spectral features, and Mel spectrogram combinations. VGG16 was chosen for the SER framework due to its practical performance and trade-off between prediction accuracy and model training time on different tested CNN-based networks, including ResNet, MobileNetV1, VGG16, VGG19, and DenseNet.

3.4.3 Method 2: VGG-optiVMD

A series of experiments were implemented to investigate VGG-optiVMD capabilities, incorporating nine distinct feature vectors, and utilizing a consistent computing environment throughout all tests (Intel Core-i7 processor, NVIDIA GT1080 GPU, 32GB RAM, Windows 10 OS). The implementation used the Keras framework and Python 3.8 to program quickly. The number of sample voices from EMODB and RAVDESS databases was expanded to ensure a robust dataset, stratifying them based on mean duration extremes. Acoustic feature extraction was executed using the Librosa tool, with frame size = 2048, HOP length = 256, and sampling rate = 88,400, where these parameters were selected to mitigate spectral leakage and boost frequency resolution. Pretrained VGG16 networks were trained on the extensive ImageNet database, containing more than 14 million images, to detect subtle variations within the feature maps presented.

The objective in deploying this framework was to enhance informative data encapsulated within feature vectors drawn from speech signals, thus improving member emotion prediction accuracy. The algorithm established initial K and α, and optimal counterparts were determined by iterative testing, aligning with configurations that ensured the highest test accuracy. Other hyperparameters from VMD remained constant. The DC parameter at zero was set at $\omega = 1$ to address inherent DC voltage offset typical in speech signals; tolerance parameter, which governs minimum update rate for ω, $tol = 10^{-9}$, with noise tolerance parameter $\tau = 0$. The VGG-optiVMD algorithm used two decomposition parameters $K(2--6)$ and $\alpha(2000--6000)$, because more computing power and space were required as more modes were decomposed.

The VGG16 architecture was set up to use the ADAM optimizer with learning rate of 0.0001. The network's six fully connected hidden layers were activated by the ReLU, SELU, and TanH functions, operating over 50 epochs with batch size = 4. The output layer utilized SoftMax, which completed the architecture design for effective emotion classification. Therefore, Parameters for VGG-optiVMD were calibrated as above to fine-tune the VGG16 architecture processing capabilities.

3.5 Results and Discussion

3.5.1 Method 1

Various evaluation metrics were used to compare between the methods, including confusion matrix, precision, recall, F1 score, and accuracy. Comparisons included different sample rates, feature map sizes, dimensionality (1D, 2D, and 3D), and impacts from varying the number of subsamples were examined by augmenting the window size and sample rate. I compared the performance outcomes for the proposed hybrid feature map extraction strategy compared with conventional approaches using ten different feature map representations.

Table 3.2 compares the proposed approach with several best-case conventional approaches for various methods and options applied. Mel spectrogram harmonic and percussive components are powerful predictors for emotion recognition, and the proposed hybrid feature map representation outperformed other well-known feature combination techniques. Model accuracy improved with increasing sample rate and window size since the feature map generator was able to process a greater volume of data points. VGG16 network outputs also benefited from more enriched feature sets and higher sample rate.

Table 3.2: Prediction accuracy impacts from feature extraction technique, sampling rate, and window size on EMODB dataset.

Window size	512	1024	2048
Sample sate	22050	44100	88200
Feature extraction methods	Acc. (%)	Acc. (%)	Acc. (%)
1D MFCCs	65.81	68.39	69.03
1D Mel spectrogram	75.48	75.48	82.71
1D chromagram	80.01	80.13	81.29
1D Tonnetz	56.77	63.08	56.81
1D spectral	54.84	50.93	47.10
2D MFCCs+chromagram	83.87	83.23	91.59
2D Mel spectrogram+MFCCs	88.39	85.16	85.81
2D Mel spectrogram+Spectral	82.01	85.13	80.65
3D Mel spectrogram+MFCCs+chromagram	83.87	88.39	81.94
2D log-MSS+Avg.HP(proposed)	92.02	89.54	**92.79**

Table 3.3 shows the model confusion matrix is proficient at recognizing anger, happiness, and fear emotions, but relatively weaker for the more subdued emotions (neutral and boredom).

Augmenting data points within the subsamples created significant memory storage requirement, extending to the gigabyte range, to store the base, training, validation, and test feature map files in pkl format. The proposed mode, in particular required almost 3 GB storage at 88 kHz signal sampling rate and window size = 2048 to analyze the complete voice files from EMO-DB, presenting a significant implementation obstacle for practical real-world applications.

Table 3.3: Confusion matrix for the proposed model, achieving 92.71% average accuracy on the EMO-DB dataset

Emotion:	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	**94.92**	0	0	0	5.12	0	0
Boredom	0	**78.77**	0	0	0	9.9	11.54
Disgust	0	0	**89.47**	0	9.8	0	0
Fear	0	0	0	**96**	0	0	3.85
Happiness	0	0	0	0	**100**	0	0
Neutral	0	12.81	0	0	0	**88.87**	0
Sadness	0	0	0	0	0	0	**100**

3.5.2 Harmonic-Percussive Mel spectrogram

This thesis explored potential Mel spectrogram components through a hybrid feature engineering approach, devising a new acoustic feature extraction method to enhance SER. A significant facet of this study involved fusing distinct elements, such as the harmonic, percussive, and log Mel spectrogram components extracted from speech signals. A specialized feature map generator function was built to create an enriched 2D feature map vector, and a CNN-based transfer learning strategy was employed to decode emotion data from extracted acoustic features. One significant drawback for CNN networks is their inability to disclose the patterns they uncover in the data. Furthermore, although high sampling rate improved accuracy, this also required substantial storage space for the feature maps, creating practical obstacles. Consequently, the 2D feature extraction strategy was a trade-off between memory usage and network performance. Despite this limitation, the proposed H.P. SER model performance was unaffected, achieving 92.79% test accuracy.

Future directions for this research will be to diversify the network architecture. Considerable potential remains to create more comprehensive models by combining outputs from various neural networks, each trained on different acoustic features. Incorporating call transcripts as a textual feature could further generalize the model, considering variations across languages and cultures. The ultimate goal is to develop a multimodal learning model that transcends just acoustic analysis, but embodies a multifaceted approach to member emotion and engagement level. The Python Keras based network implementation for the proposed model and more experimental results and visualizations are available in the GitHub repositories[1] noted below.

3.5.3 Method 2

Table 3.4 summarizes the outcomes for various mode count (K) and bandwidth control (α) using the proposed approach. There is strong correlation between K, α, and classification accuracy. Acoustic characteristics were significantly enhanced for particular decomposition parameter sets, with optimal settings $K \in [4,6]$ and $\alpha \in [2000, 4000]$. However, the ranges were restricted to $\alpha \in [1000, 100000]$ and $K \in [2,8]$, to reduce computational burden, which increases significantly for $K > 8$, particularly with sample rate = 88.4 kHz. This boundary served as a functional constraint within the VGG-optiVMD algorithm. Despite these constraints, the proposed methodology using 3D Mel Spectrogram+MFCCs+Chromagram achieved state-of-the-art outcome (96.09% with $K = 6$, and $\alpha = 2000$) on EMODB dataset. Also, we achieved the highest accuracy of 92.14 % with $K = 6$, and $\alpha = 4000$ using the spectral feature on RAVDESS dataset. Therefore, using the spectral feature extraction method outperforms on RAVDESS than EMODB dataset.

Table 3.4: Emotion classification performance: Automatic decomposition parameter (K and α) selection using VGG-optiVMD

Features		VGG-optiVMD Performance									
	Database	[K=4, α=2000]		[K=4, α=4000]		[K=6, α=2000]		[K=6, α=3000]		[K=6, α=4000]	
	EMO-RAV	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CH	EMODB	68.54	68.37	81.63	81.47	94.05	94.88	94.90	91.10	95.41	95.11
	RAVDESS	70.23	70.55	82.73	82.96	85.21	85.92	79.81	79.79	47.49	46.53
MS	EMODB	91.84	91.86	93.15	93.07	95.19	95.07	95.34	94.98	95.92	94.89
	RAVDESS	64.21	64.69	71.36	71.55	75.28	75.95	84.19	84.68	87.25	88.11
MF	EMODB	48.1	46.92	65.16	64.42	64.87	65.18	56.12	56.57	67.64	66.9
	RAVDESS	42.64	41.77	53.29	52.14	55.61	56.80	51.81	51.44	41.86	40.46
SP	EMODB	94.27	93.11	93.01	92.95	93.88	93.07	93.44	93.37	94.02	93.87
	RAVDESS	89.25	90.11	78.48	79.21	91.28	92.88	90.70	90.10	**92.14**	93.55
TZ	EMODB	74.93	75.11	91.25	90.89	88.92	88.91	91.84	91.12	92.44	92.10
	RAVDESS	48.21	48.26	51.04	51.67	52.07	52.12	49.06	49.12	51.98	52.23
MS+SP	EMODB	89.62	90.85	88.76	89.08	88.2	88.13	95.92	96.11	95.41	95.12
	RAVDESS	78.33	78.12	74.37	74.79	78.52	78.78	81.38	81.42	81.84	81.91
MF+SP	EMODB	58.1	58.2	66.91	66.98	65.16	65.11	62.54	62.13	67.64	67.21
	RAVDESS	53.08	53.12	56.25	56.68	60.28	60.94	58.21	58.14	54.7	54.06
MF+CH	EMODB	85.21	85.2	84.35	84.36	90.14	90.13	87.41	87.52	90.82	90.82
	RAVDESS	51.29	51.35	54.25	54.89	53.65	54.66	55.13	55.12	56.08	56.84
M+M+C	EMODB	86.56	86.42	87.41	87.35	**96.09**	96.04	93.54	93.42	94.73	95.98
	RAVDESS	60.28	60.11	60.28	60.84	61.55	62.36	59.25	60.87	57.70	57.56

Abbreviations
M+M+C: 3D-Mel Spectrogram+MFCCs+Chromagram;
CH: Chromagram; TZ: 1D-Tonnetz; MF: MFCCs; MS+SP: 2D-Mel Spectrogram+Spectral;
The top-performing results across both databases are highlighted in bold font.

[1] https://github.com/DavidHason/ser

3.5.4 VGG-optiVMD

The proficiency to precisely recognize emotional states solely from voice becomes very important for situations where only auditory cues are available, such as emergency call centers or customer service lines in organizations. Therefore, this thesis investigated how member emotions are sent through vocal intonations, with a view to identifying an optimal analysis pathway to enhance this aspect. Experimental outcomes confirmed that VGG-optiVMD, a version of the VMD algorithm, significantly improves SER performance. The experiments indicated that sampling rate and VGG-optiVMD decomposition parameters (K and α) are key factors determining the emotion classification system effectiveness, and combining feature vectors by concatenating significantly improved VGG network training.

However, extending the range for K and α requires caution due to substantial increases in computational demand, a constraint inherent to the VGG-optiVMD algorithm. Notwithstanding, lower decomposition parameter ranges consistently yielded higher classification accuracy. Future investigations should explore whether leveraging only the most informative decomposed modes for acoustic feature extraction could reduce computational overheads.

Potential for prediction improvement by applying the VMD algorithm upstream of the acoustic feature extraction process requires further exploration. The opportunity could revolutionize practices within human-computer interaction (HCI) and customer or member behavior analytics, enhancing our understanding and interaction with technology on an emotional level. Insights gained from this study not only propel us towards more empathetic and intuitive HCI interfaces but also provide a strategic framework for enhancing member engagement in organizations. The details of Network implementations implementation details are available on the GitHub repository[2] noted below.

3.6 Summary

This chapter empirically assessed the advantages and drawbacks for para linguistic emotion recognition methods such as H.P SER and VGG-optiVMD on many benchmark datasets, and investigated the effects from setting speech signal decomposition parameters on classification. The VGG-optiVMD approach excels primarily due to its innovative integration of VMD with the VGG network, enhancing emotion recognition in speech.

[2] https://github.com/DavidHason/VGG-optiVMD

VMD dynamically decomposes speech signals into distinct frequency modes, capturing nuanced emotional details more effectively than traditional static features. The method optimizes decomposition parameters K and α adaptively, ensuring optimal feature extraction. Coupled with the VGG16 network, renowned for its deep learning prowess, this integration allows for superior pattern recognition in emotional data, leading to more accurate and robust emotion classification compared to baseline models. Therefore, the proposed VGG-optiVMD approach was shown to be superior to previous SER algorithms, and automatic selection for optimal K and α greatly improves VGG-optiVMD performance, avoiding setting them by hand, which requires domain knowledge or expert guidance. The outcomes from this chapter provide the following key insights.

- Advanced SER techniques, such as like VGG-optiVMD, are pivotal for interpreting emotions from voice data, impacting member engagement and churn analysis.

- Speech signal processing is essential for emotion detection in voice-only contexts, such as emergency services or customer support.

- The VGG-optiVMD algorithm's state-of-the-art results on emotion classification demonstrate its potential to revolutionize member sentiment analysis.

- Future advances in SER, such as reducing computational demands while maintaining high accuracy, will enable more dynamic and responsive member engagement strategies.

CHAPTER 4

Member Financial Literacy Prediction Role in Churn Analysis

This chapter explores how financial literacy (FL) influences member churn. FL empowers informed financial decisions, impacting member engagement, satisfaction, and decision making. Low FL leads to dissatisfaction and increased churn risk. This chapter considers

1. considers FL impact on engagement,
2. considers FLs vital role as a predictive factor in member churn analysis, and
3. introduces the SMOGN-COREG model, semi-supervised regression framework to address unlabeled and imbalanced data.

4.1 Background and Motivation

Financial literacy is important for member engagement, particularly within financial institutions. Members well-informed about financial products and services are more likely to make informed decisions that align with their personal and financial goals, leading to higher satisfaction; whereas members with limited FL may feel overwhelmed by offered financial product complexity, resulting in lower engagement levels, dissatisfaction, and higher likelihood of exiting the organization. Therefore, The ability to understand financial concepts and management is crucial for individual and society economic wellbeing, particularly in the contemporary world.

Suppose we accept the hypothesis there are four basic financial domains: personal funding, savings, credit, and investment affairs. Then we can define FL as the ability to analyze and handle costs and progress in terms of financial management, i.e., high FL represents the ability to turn an asset into income. The modern definition of FL has been updated to include digital currencies, i.e., cryptocurrency, virtual currency, and central bank digital currency. However, FL cannot be summarized by the above definition, since there are many unknown factors that could make the FL definition more sophisticated among researchers and financial analysts, including sex, age, income level, education level, occupational status, demographics, geographical location, language, and ethnicity [44, 46, 115, 116].

Aside from all these above definitions, a general definition for FL has been stipulated in a report to the National Foundation in the UK for Education Research by Schagen and Lines [115] as financial literature is "the ability to make informed judgments and to take effective decisions regarding the use and management of money". This definition has been used many times [115]. There are two main qualitative and quantitative research approaches toward FL: qualitative research concerning FL definition, concept, and evolution, based on surveys; and quantitative, which focuses on predicting FL levels and impacts from low FL.

Financial literacy is not only essential for people, but is also vital for member-centric financial organizations to achieve efficiency and success in the market. It has been determined that current economic conditions have raised significant concerns about the financial security of Australians, in particular for those people who seem to lack the resources and skills required to withstand downswings in the market and take advantage of upswings. Individuals are generally responsible for several financial decisions, the two most are retirement preparation and house financing. There is a relationship between the complexity of these choices and increasing stakes. For example, current economic issues have highlighted the significance of making effective financial decisions, and also the consequences or results of making financial decisions without sufficient FL [117].

A few qualitative FL studies have been conducted by the Australian and US governments. For exaqmple, the US has realized many significant surveys aimed at determining student FL, and many other FL initiatives have been implemented throughout the world to predict FL for the various groups. For example, the Canadian Bankers Association and Enterprise New Zealand Trust implemented the same program; and Australia implemented various reports to determine if there wais a need for not only better understanding but also improving FL. Although the latter survey concluded that most

Australians tend to have fundamental FL, those from lower socioeconomic backgrounds are disadvantaged in terms of making informed decisions about money management. However, almost all of these studies rely on literature reviews and online surveys, and a few have uses data mining to predict people's FL. Most also employed only simple linear regression to determine correlations.

Many recent studies have considered FL definition and effects, some through online surveys to measure FL level, asking volunteers questions including compound interest, inflation, time value of money, and risk diversification . Risk diversification has been the most challenging question, with only 9% of respondents able to provide a correct answer in Australia [45].

Portfolio diversification encompasses various strategies to optimize the balance between risk and return.

- Allocation across multiple asset classes. This involves distributing investments among diverse categories such as equities, fixed-income securities, real estate, and commodities. Each class exhibits distinct risk and return characteristics, contributing to the overall risk management for the portfolio.

- Diversification across various sectors, industries, geographical regions, and countries. This approach mitigates risks associated with specific market segments or geographical areas. Investing in a broad range of industries and regions reduces the portfolio's exposure to sector specific or region specific economic downturns.

- Investment in companies with differing market capitalizations. This strategy includes investing in large, mid, and small-cap companies. Company size can influence its market behavior and risk profile, thus opening the company size offers an additional layer of diversification.

- Varied investment durations in income-generating assets. Diversification can also be achieved by investing in assets with different maturity periods, such as short, medium, and long-term investments. This helps manage liquidity needs and interest rate risks.

The effectiveness of diversification is quantifiable using the correlation coefficient between pairs of assets. This coefficient ranges from -1 to 1, and measures the degree to which two assets move relative to each other. Correlation coefficient = -1 indicates a perfect inverse relationship, whereas coefficient = 1 signifies a perfect positive correlation, and and coefficient = zero implies no relationship. Broadly, lower correlation between assets in a portfolio signifies better diversification, as it suggests that the assets do not move in tandem, thus potentially reducing overall portfolio risk.

A common aim for organizations is to increase their members and employees FL. However, each organization benefits differently from developing FL knowledge, and the outcomes can be used differently, depending on the particular organiztion.

- Government authorities, financial planners, and fintech companies can leverage artificial intelligence to exploit unlabeled financial network data alongside online FL surveys. Integrating technology and data analytics could significantly improve CRM, member engagement, and customer retention rates, thereby enhancing overall economic value.

- Member-centric organizations could benefit from more informed decision-making regarding investments, savings, and budget management, ultimately encouraging a more financially responsible and empowered community just from inceasing FL among students and faculty.

- Enhancing FL levels is a multidimensional goal for research and development engineers, that intersects with user experience, member engagement, system engineering, and business-oriented predictive analytics. Equipping individuals with better financial understanding through artificial intelligence, would enable engineers to design more efficient financial tools to explain complex financial products to members or users. This would assist members in making informed decisions and provides a substantial financial behavior dataset from which to optimize algorithms, ultimately contributing to more personalized and effective financial services.

Few previous studies used only supervised learning and labeled data for classification tasks. Real-value target variables in regression tasks face a practical difficulty in implementating semi-supervised regression (SSR) algorithms. SSR builds a better regressor by utilizing a large amount of unlabeled data with a small amount of labeled data; requiring less human effort while providing better performance in theory and practice. However, no previous study used semi-supervised learning (SSL) methods for FL prediction. Therefore, there is possibly considerable scope to improve FL prediction using SSR over unlabeled data.

Many studies have shown that low FL increases not only the churn risk in organizations but also the social harm risks, such as low-income retirement, job loss, mental health problems, and longevity outcomes. Reliable FL prediction algorithms are essential to estimate individual FL for allocating specific intervention programs and financial advice to less financially literate groups. This will not only increase organisational profitability and social economy, but also reduce government spending.

4.1.1 Tackling online survey challenges with semi-supervised learning

Online surveys are not always ideal to measure FL levels in a large organization, sometimes other methods, including using machine learning to analyze recorded data, are more economical and constructive. Although most of these data types data in finance networks are unlabeled, exploiting the large amount of easily accessible unlabeled data through SSL is the best strategy when collecting labeled data is too expensive or infeasible. Semi-supervised learning aims to find meaningful features from an unlabeled dataset and use them in the prediction model. The main reason why SSL has recently become a popular approach is because the amount of unlabeled data is increasing very quickly.

This thesis proposed a novel method to synergize a limited dataset from a qualitative online FL survey, including diverse queries indicative of members' FL. Survey findings assign a numerical FL value, ranging from 0 to 1, to each participant, and this FL dataset is subsequently utilized within an SSL framework to categorize a substantial amount of unlabeled data. This process, termed "financial X-Ray", is designed to provide detailed analyses of members' financial engagements within an organization, by combining qualitative data from surveys with quantitative SSL methodologies to assess member FL levels.

4.1.2 Key contributions

The main contributions from this thesis are summarized below.

- Developed the SMOGN-COREG model, an innovative semi-supervised regression framework that effectively leverages unbalanced and unlabeled financial datasets.

- The proposed approach handles unbalanced datasets by merging oversampling strategies with co-regression algorithms, enhancing the predictive power of a combination of labeled and unlabeled data.

- First known use of SSL for financial literacy prediction, a notable breakthrough for utilizing vast unlabeled datasets in a domain where such methodologies have not been extensively applied.

- The proposed model achieved enhanced accuracy with synthetic samples; labelling 64% of prior unlabeled data.

4.2 Preliminary Knowledge

4.2.1 Semi-supervised learning

Low cost and broad access to unlabeled data in various research scopes has made the SSL method more popular. On the other hand, the recent massive growth of unlabeled data on different platforms, such as social media, education systems and finance networks, makes it inevitable to ignore them in real-world predictive models. SSL has been categorized into two main approaches: semi-supervised classification (SSC) and semi-supervised regression (SSR) based on the target variable type in model output. SSC is used where the target variable is discrete, whereas SSR is the best choice when the model output is a continuous variable. This thesis focuses on SSR since FL is a continuous variable. SSL exploits both labeled and unlabeled data to obtain higher accuracy, where most data is unlabeled and only a small amount of labeled data is available. The underlying method can be summarized as

$$X = (x_i)_{i \in [n]},$$

where n there is the total number of instances, x is an independent predictor into labeled set $X_l = (x_1, ..., x_l)$ associated with labeled data $Y_l = (y_1, ..., y_l)$, and unlabeled instances $X_u = (x_l + 1, ..., x_l +_u)$ where labeled data are not available for them. The SSR algorithms can be classified based on the relationship between attributes in SSL, i.e., parametric or non-parametric methods.

- Parametric methods use a functional form to define relationships among attributes, such as linear, quadratic, or periodic relationship functions.

- Non-parametric methods extract relationship between attributes from input data using an unknown estimator function. Some well-known non-parametric regressors are k-nearest neighbors, polynomial estimators, and kernels.

Learner view in SSL refers to the concept that each example x can be interpreted in various aspects related to feature sets. The learning process can be categorized into single and multiple views.

- Single view is mostly used for real-world problems where each predictor describes the feature in one view.
- Multiple view is used where different views on the feature sets are considered for each instance x. The most popular multi-view learners are the genetic and random split algorithms.

There are also two types of learners: single and multiple learners. The multiple-learner method applies more learners during learning to improve prediction accuracy and reduce possible over-fitting, whereas single learners are mainly utilized in simple regression tasks. Other important factors to consider when choosing suitable SSR algorithms are the number of instances, unlabeled set pool size, number of iterations, instance confidence measurement accuracy, and evaluation metrics. Table 4.1 shows how SSR models can be categorized.

	Parametric		**Non-Parametric**	
	Single View	**Multi-view**	**Single View**	**Multiple View**
Single Learner	Least squares Regression	-	Simplified Co-Regression Kernel Ridge Regression SVM Regression Output Kernel Regression Graph Laplacian Regularisation Hessian Regularization Parallel Field Regularization Spectral Regression Local Linear Regression Gaussian Process Regression Hybrids	Simplified Co-Regression
Multiple Learners	-	-	Co-regression	Kernel Ridge Regression Hybrids

Table 4.1: Summary of Regression Techniques

Semi-supervised learning uses unlabeled and labeled data in the learning process, in contrast with supervised and unsupervised learning methods. Labeled data are expensive and/or time-consuming to obtain and require experienced human annotators. In contrast, unlabeled data acquisition is relatively easy and can quickly obtain considerable data for the learning process. Although exploiting unlabeled data via SSL helps to reduce human effort and improve model performance, there are other challenges that make a time-consuming effort in model tuning more critical than other machine learning techniques. This thesis designs a regressor that utilizes unlabeled data in the training set to perform better than a regressor that uses only labeled training data [118].

Commonly employed SSL methods to fit the problem structure with generative mixture models include self-training, co-training, transductive support vector machines, and graph based methods. An assumption is required to design a new algorithm when the current SSR method is hard and complicated to modify. Inductive semi-supervised and transductive semi-supervised learning vary depending on the model application. Inductive semi-supervised learning can handle unseen data, whereas transductive only works on the existing labeled [119]. The problem in modeling a SSL can be formulized in labaled

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{l} \overset{iid}{\sim} p(\mathbf{x}, y)$$

and unlabeled

$$\{\mathrm{x}_i\}_{i=l+1}^{l+u} \overset{id}{\sim} p(\mathbf{x})$$

training data, where L and U are labeled and unlabeled data, respectively; X is an input data point, y is a target label, $P(X, y)$ is the unknown joint distribution, and $p(X)$ is marginal (typically $p(X) = l \ll u$). The transductive method only considers labeled data [119]

$$\{\mathbf{x}_i\}_{i=l+1}^{l+u}.$$

Thus, the proposed SSL method can be expressed as

$$X = (x_i)_{i\in[n]},$$

where n is a total number of instances and x is an independent predictor into the labeled set $X_l = (x_1, ..., x_l)$ associated with labeled data $Y_l = (y_1, ..., y_l)$ and unlabeled instances $X_u = (x_l + 1, ..., x_l +_u)$, where labeled data are not available.

4.2.2 GOREG for unlabeled data

Semi-supervised learning uses unlabeled and labeled data in the learning process, in contrast to supervised and unsupervised learning methods. A non-parametric multi-view SSR method is the most flexible algorithm for several domains, therefore, this thesis proposes a non-parametric multi-learner SSR algorithm inspired by co-training SSR. The co-training algorithm is mostly used for classification, where the algorithm trains two supervised learning classifiers separately on independent sufficient and redundant view sets, and both are independently applied to the classifiers to predict unlabeled examples to augment the training set. This algorithm is used in many fields, including statistical analysis and noun phrase identification. A better SSR algorithm for labeling examples shouild have low computational load and provide superior results in a large, imbalanced dataset. The co-training regressors (COREG), is a SSL algorithm that implements two regressors, one of which labels unlabeled data for the other, and the confidence in labeling an unlabeled example is determined by the sum of the mean squared error reduction over the labeled neighborhood for that example. Final predictions are derived by averaging the regression estimates generated by both regressors. COREG utilizes different distance metrics rather than requiring sufficient and redundant views, hence it has broad applicability. COREG employs a lazy learning method including two k-NNs in the learning process, and can improve computational load since it doesn't hold a separate training phase and optimizes regressors in each iteration. Whereas, the large number of labeling iterations required for neural networks or regression trees will lead to heavy computational costs [120].

COREG employs the k-NNs to compute the MSE for each X_u and hence identify the most confidently labeled example by maximising $\Delta_{\mathbf{x}_u}$ in

$$\Delta_{\mathbf{x}_u} = \sum_{\mathbf{x}_i \in \Omega} \left((\mathbf{y}_i - h(\mathbf{x}_i))^2 - \left(\mathbf{y}_i - h'(\mathbf{x}_i)\right)^2 \right), \tag{4.1}$$

where h and $h^{'}$ are the original and refined k-NN regressor, respectively; and Ω is the set of k-NN labeled examples of X_U. Information provided for regressors $(\mathbf{x}_u, \hat{\mathbf{y}}_u)$ by where $\hat{\mathbf{y}}_u$ defines with $\hat{\mathbf{y}}_u = h(\mathbf{x}_u)$.

Following the co-training method, both base regressors are first trained on the primarily labeled set, where its size is set to the labeled ratio (R),

$$R = \frac{|L|}{|D|}, \tag{4.2}$$

where D is the primary training set, L the target examples, and U the initial unlabeled set, such that $D = L \cup U$ and $|L| \ll |U|$ and the parameter defines the ratio between the L training set size and total number of examples.

4.2.3 SMOGN for imbalanced dataset

SMOGN integrates random undersampling with two oversampling techniques to enhance generated data diversity through Gaussian noise. This technique creates new synthetic instances utilizing SMOTER, which identifies k-NN by measuring distances between data points and can also add Gaussian noise to further diversify the data [121, 122]. SMOTER determines the neighborhood proximity, categorizing them into 'safe' or 'unsafe' zones by evaluating median distance between data pairs. The approach systematically segregates important from less important cases into respective BinsR and BinsN partitions, and finally applies oversampling and random undersampling strategies. This thesis applied SMOGN sampling on all datasets to balance target variable distributions.

The unbalanced learning problem is concerned with learning algorithm performance in the presence of underrepresented data and severely skewed class distributions [123]. We can solve this skewness by defining a relevance function to determine normal and rare value sets, and then map them onto a relevance scale between 0 and 1, representing minimum and maximum relevance, respectively [124]. A threshold t_R was established on relevant values assigned to each user to define the rare value set as

$$D_R = \left\{ |x, y| \in D : \phi(y) \geq t_R \right\},$$

and normal cases as

$$D_N = \left\{ \langle x, y| \in D : \phi(y) < t_R \right\},$$

where

$$\mathscr{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$$

is a training set with N data points. The relevance function and t_R are used to determine D_R and D_N sets in all sampling strategies.

Figure 4.1 shows that SMOGN's synthetic graphics for baseline cases consider five proximate neighbors, three of which fall within the defined safe zone, and the remaing pair lie beyond, in the unsafe zone. This illustration highlights that normal bin instances (marked in green) are more likely to intersect with significant bin instances within the unsafe zones. Therefore, SMOGN synthesizes new instances, with SMOTER choosing either k-NN or Gaussian noise as the mechanism for new instance generation, contingent upon the spatial relationship between data points. Interpolation through SMOTER is used when a neighbor is in a safe zone, and Gaussian noise is used when a neighbor is in an unsafe zone.

4.3 Methodology

This section explores the SMOGN-COREG model, a proposed technique for SSR learning. This model is particularly useful when labeling unlabeled data, since this requires an algorithm that can handle large unbalanced datasets efficiently with low computational load while delivering high quality results. Therefore, selecting an SSR algorithm is

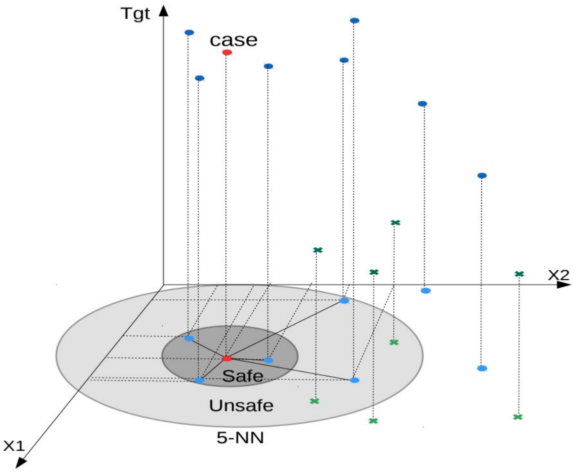


Figure 4.1: Synthetic examples in SMOGN

pivotal, and the SMOGN-COREG model is proposed as an effective solution for these requirements.

4.3.1 SMOGN-COREG

Practical study, not detailed here, empirically investigating different sampling and SSR algorithm adaptability confirmed that combining sampling strategies with a non-parametric multi learner SSR algorithm achieves superior results on an imbalanced dataset. The best arrangement was to combine compatible SMOGN and COREG in a real-world financial dataset. Rare but important data points in the minority class can often create bias in SSL models. Therefore, this study employed the SMOGN method as a pre-processing step to improve model performance during the learning phase. This technique strategically combines undersampling overrepresented values typically clustered around the normal distribution mean of the response variable and oversampling underrepresented, or 'rare', values located in the distribution's tails. SMOGN enhances the regression process by introducing Gaussian noise to modify the interpolated values synthetically. This process involves applying a function ϕ to the response variable, which assigns a corresponding variate $\phi \in [0,1]$ for each data point, determining its classification as a majority or minority instance based on the predefined threshold t_R.

SMOGN randomly selects from the observed values within their function range to generate synthetic values for categorical features. The resulting post-processed data frame includes a balanced mix of under and oversampled observations. These sampling strategies are crucial as they improve the learning process by enhancing representation for rare but critical cases. Figure 4.2 shows that the workflow to implement the proposed SMOGN-COREG model involves four primary stages: input data; pre-processing (including data cleaning), feature selection, and sampling; follwed by data augmentation of the labeled set using SSR, and increased labeled data in output.

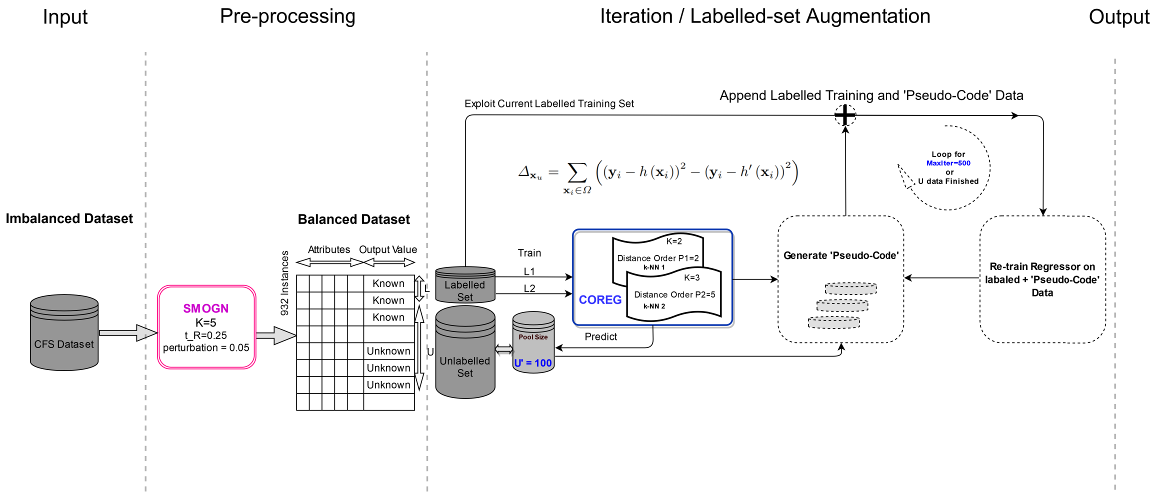


Figure 4.2: Proposed SMOGN-COREG model workflow

This thesis utilized the COREG algorithm due to its ability to work with diverse distance metrics, eliminating the requirement for numerous redundant data views. COREG incorporates sampling strategies with a non-parametric, multi-learner, SSR algorithm, to significantly enhance model performance, particularly when dealing with unbalanced datasets. Base regressors undergo co-training on a predominantly labeled dataset, which is determined by the ratio R for labeled set L to the entire dataset D, which also includes the unlabeled set U.

4.4 Experiment Outcomes

4.4.1 Materials

Table 4.2 summarizes the five imbalanced datasets that experiments were conducted on. One dataset, includes 68 feature (55 integer and 14 real variables) and 932 instances, was provided from an online survey in 2017 and 2018 by a local Australian superannuation company with more than 1 million customers; and the other four datasets with 89 features (54 integer, 16 polynomial, and 19 real variables) and 918 labeled instances belong to members who participated in the FL survey. Thus the combined dataset features cover customer financial activities, demographics, financial status, income, account balance, marital status, age, and employment.

Table 4.2: Datasets

Dataset	# Attributes	# Instances	Records
CFS_201706	89	824	73336
CFS_201712	89	856	76184
CFS_201806	89	899	80011
CFS_201812	89	918	81702
CFS_2017-2018_FL	69	931	64239

4.4.2 Baseline model

The baseline model for this comparison was the meta multi-scheme SSR Algorithm (MSSRA), proposed by Fazakis et al. [125]. This algorithm incorporates three base k-NN regressors, with k = 3, 7, and 9. The approach begins with these k-NN models to label the data, which is then followed by self-training to enhances the initial labeled dataset by incorporating insights drawn from the unlabeled data set.

The random forest (RF) regressor was employed to retrain the model and integrate refined labels obtained from the previous iteration. The iterative process completed with the RF regressor outputting labels for the test instances, which were previously unknown. A unique aspect of MSSRA is that it leverages different regressors within and outside the iterative learning process. This multi-regressor approach effectively introduces diverse perspectives into the model training process, and this diversity in the learning process contributes significantly to the resulting model robustness, enhancing its overall predictive performance.

Table 4.3 shows the suite of supervised and semi-supervised regressor models (built on the Weka platform) compared with the proposed SMOGN-COREG model performance. Several additional supervised regressors were also investigated, but are not discussed explicitly since their performance was markedly inferior.

4.4.3 Experiment setup

First, a cross-validation method was implemented, dividing the datasets into 10 folds. One fold was reserved for testing and the remaining folds were allocated for training. The unlabeled ratio, UR= 80%, was used to split the training set for each fold, retaining only 20% of labeled data for the learning process. COREG maximum iterations = 100,

Table 4.3: Regressor models for comparision

Method	Characteristics	Parameters
LR	Linear relationship	Weighted instances
k-NN	Euclidian distance	$k \in \{4,7,9\}$
SMOreg	SVM-polynomial kernel	Batch size = 100
M5 Rules	Model tree in if-then form	Min. instances/leaf = 4
M5 Model Tree	Multivariate LR trees	Min. instances/leaf = 4
Random Forest	Regression/classification	Depth: unlimited; iterations = 100
MSSRA(Baseline)	Semi-Supervised	Base regressors: 3,7,9 k-NN, RF

with pool size $U^{'}$ = 100, ensuring that $\Delta_{x_u} > 0$ always in each iteration. This configuration resulted in a theoretical maximum labeling capacity of 50,000 iterations. However this number seemed overly optimistic, considering potential negative effects from noisy data in the L subset. After experimenting with the trade-off between iterations and model runtime, 500 iterations with pool size $U^{'} = 100$ unlabeled data points was identified as ideal, ensuring all confidence predictions to enhance the labeled set during learning across all five datasets.

The two k-NN regressors within COREG employed the selected distance order $k = 2$ and 3, respectively, and $k = 2$ for the SMOGN algorithm oversampling, with $t_R = 0.25$. Gaussian noise included in the SMOGN was set at 5%, producing 0.05 perturbation 05 and maximum iteration count = 1000. The pool comprised 100 unlabeled examples randomly chosen from the unlabeled set in each iteration, with the final prediction being the averaged regression predictions from both regressors. Average MSE for labeling was also recorded for the most confident instances.

The proposed approach was benchmarked against one SSR algorithm and five commonly employed supervised regressors across five distinct datasets (see Table 4.2). multivariate linear model defined as

$$Y_i = \beta_0 + \sum_{j=1}^{P} \beta_j X_{i,j} + \varepsilon_i,$$

where Y_i and $y_i^{'}$ are the actual and predicted values, respectively.

Four evaluation metrics were considered to evaluate regression performance, as shown below. Optimal predictions are indicated by higher PCC and R^2 values and lower MAE and $RMSE$ values.

Root mean squared error (RMSE) is a standard metric measuring the difference between predicted and actual values,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}, \tag{4.3}$$

where RMSE = zero indicates perfect predictions with no deviations. However, MSE can disproportionately reflect impacts from larger errors, complicating interpreting model performance.

Mean absolute error (MAE) measures predictive accuracy considering absolute error values,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i'\right|, \tag{4.4}$$

facilitating a more straightforward interpretation compared with RMSE. Similar to RMSE, MAE is always non-negative, and lower values denote more precise fits, and MAE = zero signifies roboust predictions.

Coefficient of determination (R^2) commonly referred to as R^2, determines the proportion of variance in the observed data that the model accounts for. It is calculated using the ratio of the sum of squared errors (SSE) from the predicted mean y_i' to the actual observations y_i, relative to the total sum of squared errors (TSS) between the actual observations y_i and overall mean of them $\overline{Y}_i$,

$$R^2 = 1 - \frac{SSE}{TSS} = \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \overline{Y}_i)^2}, \tag{4.5}$$

Typically $R^2 \in [0,1]$, where $R^2 = 0$ indicates the model fails to capture any variance within the dataset, and $R^2 = 1$ indicates a perfect fit, describing the entire data variance.

Pearson correlation coefficient (PCC) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables,

$$PCC = \frac{\sum_{i=1}^{n}(y_i - y_i)\left(y_i' - y_i'\right)}{\sqrt{\sum_{i=1}^{n}(y_i - y_i)^2}\sqrt{\sum_{i=1}^{n}\left(y_i' - y_i'\right)^2}}. \tag{4.6}$$

4.5 Results and Discussion

Considering the various experiments are conducted, the proposed SMOGN-COREG model has a significantly improves perfrmance compare with the alternative regression methodologies. One core objective was to utilize SMOGN potential to address imbalanced regression tasks. Figure 4.3 shows the empirical analysis for the FL dataset, emphasized

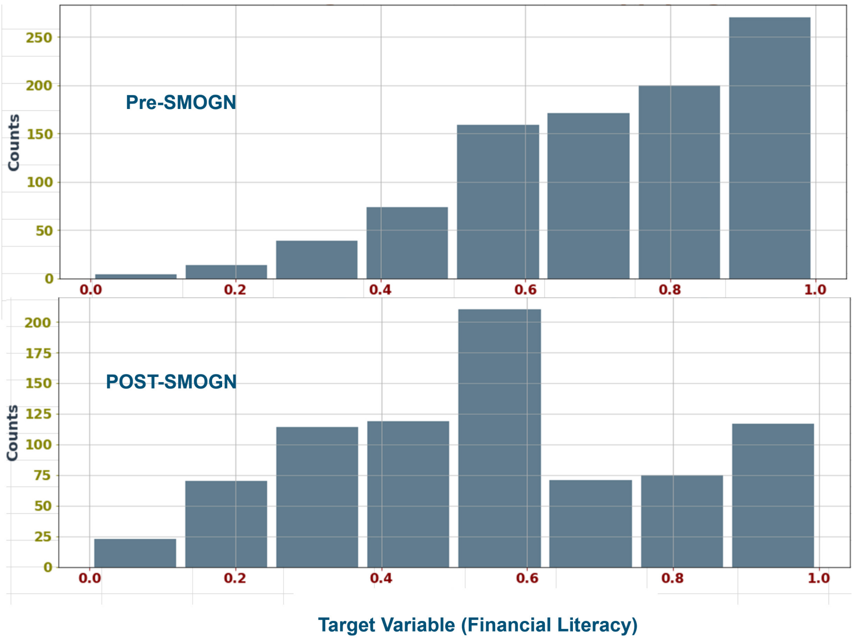


Figure 4.3: Target variable distribution pre and post SMOGN application

the SMOGN efficacy due to the significant reformation in the skewed data distribution before applying SMOGN. Figure 4.3 shows the initial imbalance in the target variable, with the initial data points densely populated above 0.5, whereas significant redistribution after SMOGN, reducing data points from 0.6 to 0.9 and considerably increase below 0.5, confirming SMOGN sampling efficacy to rectify data imbalances.

Figure 4.4 shows MAE metrics results for the datasets considered. The M5 and M5 Rules algorithms yield considerably lower MAE compared with their supervised and semi-supervised counterparts. Although the RF model achieves the least variation, MAE oscillating narrowly between 0.102 and 0.104, the superior result is achieved by the SMOGN-COREG model, which consistently records the lowest MAE = 0.099 – 0.1091 across the considered datasets.

Table 4.4 shows RMSE outcomes for the considered datasets are satisfactory for all regressors. The M5 model achieved optimal RMSE = 0.1207, maintaining the least variance across diverse datasets. The proposed SMOGN-COREG model exhibits enhanced efficacy relative to the baseline model and the three k-NN regressors, but marginally higher RMSE than M5, SMOreg, and LR. The latter result is because SMOGN-COREG regression models were trained on a broader training set, combining the initial labeled set with a pseudocode subset. This integration inherently increases prediction error in SSL relative to supervised methods due to SSL vulnerability to data noise, which can

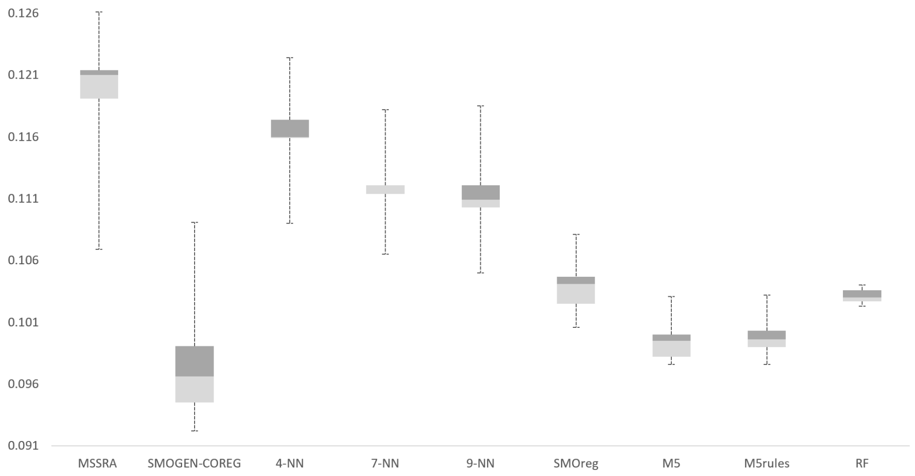


Figure 4.4: The SMOGN-COREG and M5 models achieved the better MAE result.

produce inaccurate pseudo-labels and consequently highly confident but erroneous predictions. Nevertheless, the SMOGN-COREG algorithm demonstrates superior stability over the baseline SSL algorithm and the other supervised learning models.

Table 4.4: RMSE Results

Datasets	MSSRA	SMOGEN-COREG	4-NN	7-NN	9-NN	SMOreg	LR	M5	M5rules	RF
CFS_2017-2018_FL	0.1367	0.1356	0.1344	0.1306	0.1284	0.1317	**0.1275**	0.1276	0.1277	0.1317
CFS_201812	0.1565	0.1335	0.1483	0.1439	0.1426	**0.1321**	0.1224	0.1214	0.1215	0.1339
CFS_201806	0.1618	0.1303	0.1581	0.153	0.1533	0.1325	0.1304	**0.1223**	0.1229	0.1359
CFS_201712	0.156	0.1285	0.1502	0.1447	0.1448	0.1831	1.1528	**0.1227**	0.1231	0.1362
CFS_201706	0.1549	0.1416	0.1513	0.1465	0.1462	0.1263	0.1251	**0.1207**	0.1208	0.1361

The significant improvements in the coefficient of determination (R-squared) and Pearson correlation coefficient (PCC) with the introduction of my model are indicative of the synergistic integration of the SMOGN sampling technique with my model. The results shown in Figure 4.5 and Table 4.5, reveal that the COREG algorithm alone obtained modest R-squared and PCC values of 0.4431 and 0.6656, respectively. However, the SMOGN-COREG model significantly elevates the improvement to 0.7171 for R-squared and 0.8468 for PCC. This not only confirms the synergistic compatibility of SMOGN and COREG within the financial network's dataset but also emphasizes the critical role of the SMOGN sampling technique in increasing prediction accuracy in imbalanced datasets.

Table 4.5: PCC results

Dataset	MSSRA (Baseline model)	SMOGEN-COREG	4-NN	7-NN	9-NN	SMOreg	LR	M5	M5rules	RF	Improved %
CFS_2017-2018_FL	0.7922	**0.8468**	0.7876	0.7988	0.806	0.7955	0.8092	0.8094	0.8092	0.7985	4.6
CFS_201812	0.7465	**0.8384**	0.7274	0.7472	0.754	0.7919	0.8249	0.827	0.8267	0.8119	1.4
CFS_201806	0.7322	**0.8622**	0.6837	0.7117	0.7144	0.7917	0.8025	0.8259	0.824	0.8081	4.4
CFS_201712	0.7476	**0.8523**	0.7171	0.7442	0.7461	0.6295	0.0858	0.8225	0.8214	0.809	3.6
CFS_201706	0.7501	0.7454	0.7095	0.7325	0.7365	0.8085	0.8134	**0.8268**	0.8265	0.8053	−8.8

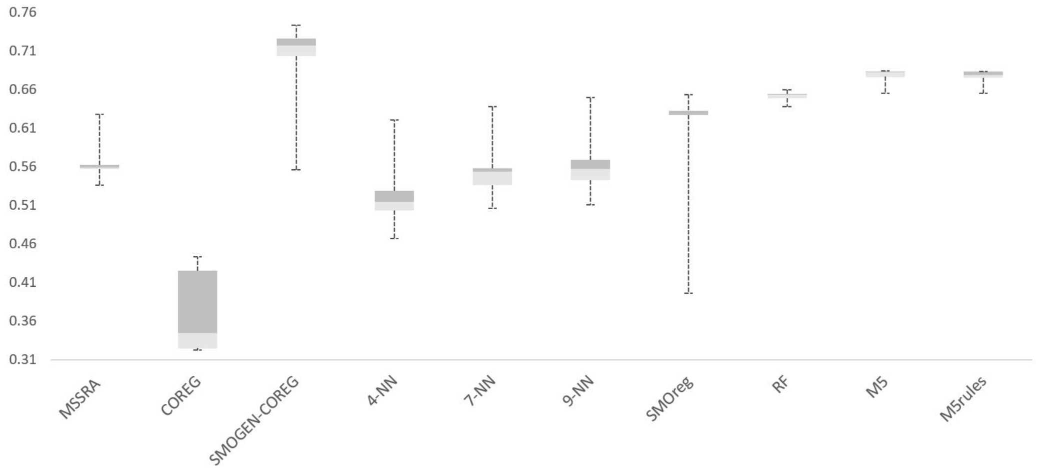


Figure 4.5: R^2 results

4.5.1 Discussion

This investigation into FL predictors for churn analysis has highlighted the critical importance of integrating financial education into member engagement and retention strategies. Financial literacy is fundamental to empower members to make informed decisions that align with their long-term financial goals, enhancing their engagement with financial services, and increasing their propensity for continued membership. The complex FL landscape extends beyond a specified understanding of financial products and services. This study is the first to use SSL to predict FL levels utilizing extensive datasets of unlabeled financial information, due to the impracticality of traditional approaches, such as online surveys, for gauging FL across large population groups. SSL is an effective strategy that leverages the abundant unlabeled datasets, extracting pertinent features that can significantly inform and enhance predictive model accuracy. The proposed semi-supervised SMOGN-COREG model represents a significant advance in addressing inherent obstacles posed by imbalanced datasets, which are common in financial contexts.

The SMOGN-COREG model was empirically shown to be superior predicting FL levels than traditional methods. The model uses advanced sampling methods to balance data, and hence improve prediction accuracy. However, this study has some limitations.

Although SSL is advantageous for processing extensive datasets, it also introduces potential complications due to noise in the data, which can result in generating inaccurate pseudo-labels. This challenge necessitates unbiased model training and validation to ensure outcome reliability. Additional data sources, such as behavioral and transactional records, should be explored to improve SMOGN-COREG models. Integrating SSL with other advanced machine learning models presents an opportunity to enhance the FL prediction field and general financial behavior analysis. The main objective is to develop a framework that provides multifaceted insight into member financial behaviors and hence improve member engagement in organizations.

The Python implementation for the proposed SMOGN-COREG model and more result visualizations can be downloaded from the GitHub repository[1] shown in the footnote.

4.6 Summary

This chapter presented a comprehensive study on financial literacy (FL) as a predictor of member churn, highlighting FL significance in enhancing member engagement and retention. A novel semi-supervised learning model, SMOGN-COREG, was proposed to predict FL using unlabeled financial data, and confirmed to be superior to traditional survey methods. This model effectively addressed dataset imbalances, demonstrating excellent robustness and accuracy. These findings emphasize the need for advanced learning models in predicting FL and suggest future directions for integrating diverse data sources to improve FL assessments and reduce churn.

The specific outcomes from this chapter can be summarized as follows.

1. Proposed SMOGN-COREG SSL as a transformative tool for FL prediction, addressing limitations for expensive questionnaire survey methods to measure FL.

2. Confirmed the proposed SMOGN-COREG model efficacy in leveraging unlabeled financial network data for predicting FL.

[1] https://github.com/DavidHason/predicting-financial-literacy

CHAPTER 5

MEMBER CHURN CAUSAL ANALYSIS IN MULTIMODAL FUSION LEARNING

This chapter addresses the underlying causal factors for member churn and provides a strategic framework for organizations to enhance member engagement, satisfaction, and long-term loyalty. Specifically, this chapter considers the following main aspects for causal analysis of member churn.

1. Comprehensive member attrition analysis by integrating predictive and causal models.
2. Identify causal factors that influence member churn.
3. Improve churn predictive models through restrained high dimensional feature space effects.

5.1 Background and Motivation

Member engagement and churn are two sides of the same coin. Churn is typically low when members are highly engaged, and churn tends to increase when engagement drops. Studying churn indirectly studies the factors contributing to or detracting from member engagement. Given that member engagement and churn are intrinsically linked, this thesis investigates churn intricacies using a multimodal approach integrating predictive and casual models. Two strategies are commonly employed by subscriber-based organizations exist to improve market share: acquiring and retaining new customers. The latter challenge is referred to as preventing churn. It is usually more expensive to attract a new member or client than to retain an existing member or client, hence investing in member loyalty is smart. It is even smarter to target that investment into at risk members or clients, rather all of them, and particularly focus on those where there is a chance to keep. Given the at risk members and/or clients are identified, organizations then need to know how to keep them. Thus, there are countless aspects to consider for churn reduction or prevention, constituting a sophisticated challenge.

Customer service embodies two member types: those who remain loyal and use the organization's services, and those who might switch to other services or stop using the organization's services altogether. These latter members are likely to leave. This creates a dichotomy: loyalists versus churners. A major objective for this thesis is to distinguish these segments clearly. The primary task involves transforming the member churn concept into a classifiable issue, subsequently addressed by deploying a data mining algorithm for predicting churn. The second but equally important task is to identify hidden causal variables, covariates, and confounders using the proposed causal model, ensuring their relevance as a cause of churn, since predictive models may not clearly reveal these causal details.

I present a framework that conducts causal churn analysis for a local financial institution, with the aim to examine this data over a 12-month period, and subsequently forecast churn for the forthcoming six months. This timeframe facilitated extracting latent factors contributing to churn. The proposed churn model integrates Bayesian networks to describe deliberate churn causation, and a novel causality analysis methodology to test hypotheses on features with high predictive power identified by Shapley additive explanations (SHAP) and partial dependence plot (PDP) analysis and can help improve causal model outcomes.

Another aspect is assessing deep feedforward neural networks (DFF NNs) for predicting churn from large sparse datasets prevalent in financial sectors. These datasets are often generated from member-centric organizations, such as associations and insurers, utilizing interval-based features in CRM systems. Recursive feature elimination (RFE) techniques were employed to manage the high dimensional data, and the outcomes were compared with ensemble ANNs and other classifiers.

Most causal churn inference studies in telecommunications, gaming, and financial sectors involve counterfactual reasoning and causal Bayesian networks [59, 60, 66]. Despite this, few studies have employed propensity score matching (PSM) [126] coupled with DoWhy [127] to explore causality [128]. Although most previous studies focused on churn prediction, few studies link causal analysis with customer churn, specifically in the financial sector. This gap inspired my investigation for a different approach to causal analysis of member churn using deep learning and PSM/DoWhy.

5.1.1 Key contributions

Key contributions toward causal analysis of churn from this thesis can be summarized as follows.

- This thesis is the only empirical investigation of causal Bayesian networks with PSM/DoWhy into causal effect impact on churn in high-dimensional sparse datasets.
- This thesis integrated different approaches, including RFE, SMOTE sampling, and ensemble ANN to address high-sparsity datasets.

5.2 Preliminary Knowledge

5.2.1 Membership churn

Membership churn refers to termination of a business relationship or reduced member engagement over a specific timeframe. Two primary marketing strategies are essential for increasing market share: acquiring new members and retaining existing members. Member acquisition costs are significantly higher than member retention costs, hance focusing on churn risk members is a wise investment.

Member-centric businesses rely heavily on retaining satisfied members, which contributes significantly to their revenue in a competitive market. Although acquiring new clients is crucial initially, retaining existing clients gradually become equally or more important. Many previous studies have emphasized retention rate impacts on the market [67]. However, members typically exhibit warning signs before churning, prompting developing churn prediction systems that focus on member behavior to identify potential churners and reasons for their churn. These factors aid in formulating efficient retention strategies, enhancing customer lifetime value and augmenting the company market value. For example, member-centric organizations often lose members with fewer interactions over time, resulting in account closure.

Figure 5.1 shows different churn categories, including voluntary churn, where members choose to leave due to dissatisfaction or unmet expectations; involuntary churn, due

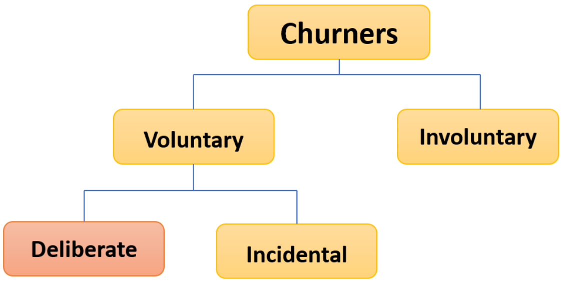


Figure 5.1: Churn categories

to payment issues or technical problems; and incidental churn due to external factors, such as location or career changes. Deliberate churn occurs when members seek alternatives due to various issues, including poor service quality, non-competitive pricing, or old technology. Organizations can utilize churn modeling to rank churn risk and consequently improve retention strategies.

The most significant reasons for churn analysis, particularly for member-centric organizations, can be summarized as follows.

- Costs to acquire a new member can be 5 times (or more) the cost to keep an existing member [67].

- Loyal members are less costly to serve, resulting in higher profits, and potentially generating new referrals.

- The loss of a member typically results in reduced profit for the organization. A company can effectively reduce member loss and increase revenue by thoroughly analyzing churn.

- Churn analysis can mitigate frustration in the business workflow. [67].

5.2.2 Calculating member churn rate

Member churn rate represents the business pulse and can be quantified in multiple ways, such as total number of customers lost, percentage of customers lost relative to overall customer base, value of lost recurring business, the percentage of dormant accounts. For example, an investing company with 400 investors that loses 8 investors in a month has 2% churn rate. Some organizations may calculate churn over quarters or financial years, but the most common method involves dividing the number of customers lost during a

Table 5.1: Churn rate calculation

Formula	Title	Description
[1]	members at the beginning of month	
[2]	Existing members who churned by the end of month	
[3]	New members in the month (not included at the beginning of the month)	
[4]	New members who churned	
[5]	Total churners	[2] + [4]
[6]	Total members at the end of month (total active)	
[7]	Basic churn rate	[5] / [6]

specific period by the total customer count at beginning of the period. Hence churn rate [55] can be expressed as

$$ChurnRate = \frac{LostMembers}{InitialMembers}. \tag{5.1}$$

The preferred churn calculation method should be clarified before finding implementable ways to deal with churn rate, as a benchmark of where the business stands and any red flag metrics. Table 3 shows member churn based on existing members who churned at the end of the month, new members who churned, and total active members at the end of the month. Therefore, churn rate can be calculated by dividing the number of churners from the beginning and end of the month by total active members at the end of the month.

5.2.3 Recency, frequency, and monetary analysis in churn

A CRM database allows combining data from static and dynamic features. Recency, frequency, and monetary (RFM) values have been utilized for many years to segment customers for churn analysis [129]. RFM analysis is a data mining model that differentiates clients by three variables, as follows.

1. Recency of engagement: the time period between the last interaction, contact, or login and present.
2. Frequency of engagement: the different number of logins to portal, calls, emails, or any member interactions in a specific period.
3. Monetary: the account's monetary value, purchased products, investments, etc.

Member behaviors can also be analyzed using the RFM method, where recency represents how recently a member has used services, frequency represents that how often members used services, and monetary refers to how much a member spent on services. RFM analysis can identify loyal members, members willing to churn, and level of member engagement.

5.2.4 Causal analysis fundamentals

Predicted data does not speak alone. Predicting churn without causality based interpretation is insufficient; it is imperative that the root cause of churn is known to formulate valid churn strategy plans. Where regression analysis predicts outcomes from variables, causal analysis reveals the root causes behind events. Variables are evaluated for their impact on results, and it seeks to confirm whether a variable influences the outcome and measure this effect. Variables that directly impact outcomes are considered causal, whereas those that move with outcomes but are not the cause are deemed correlational. It is crucial to identify those features directly leading to member churn.

Deep learning model output interpretation is divided into causal feature learning and counterfactual causal analysis. Causal feature learning involves identifying features or variables within the data that have causal relationships with the outcome of interest, whereas counterfactual explanations are a way to understand causal relationships considering alternative realities by modifying some variables and observing the hypothetical outcomes. This thesis employed causal feature learning, which is explained in detail in the following sections.

Causal analysis employs experimental design and statistical reasoning to uncover causal connections. It requires a time-ordered sequence where causes precede effects and a credible mechanism for causal influence. It also involves discovering the actual cause for a phenomenon by excluding other possible causes. The scope ranges from hypothetical scenarios to factual event evaluations, where we anticipate possible outcomes. For example, a hypothetical causal analysis might look at the effects of a price increase, while factual analyses might look at the effects of important historical events. The choice to focus on causes or effects depends on the research intent and subject. The four primary patterns underpin causal investigation, include multiple causes leading to a single outcome, a single cause with multiple outcomes, hypothetical effects that predict future outcomes, and causal chains that link events to outcomes. These patterns guide the investigation structure, ensuring a methodical approach towards causal networks. In this analytical framework, cause types are classified as immediate causes, which directly cause the effect; remote or background causes, from a more distant past; or perpetuating and hidden causes for holistic understanding.

Bayesian causal analysis uses probability to articulate uncertainties around unknown quantities based on known data. For example, it can help to quantify the likelihood that various features contribute to high churn rate, which improves our understanding of the underlying causes. Bayesian inference is the process of refining beliefs with incoming information. New information becoming available changes our ideas' likelihood and helps us make better predictions.

The core concept for causal inference is to evaluate an effect estimation using causal discovery tools such as DoWhy [127], EconML [130], Causal Discovery Toolbox [131], and TIGRAMITE [132]. The DoWhy causal inference tool can quantify effects from a predefined treatment set on an attribute, and constitutes a powerful tool for treatment effect estimation of individual feature observations. Figure 5.2 shows that the main causal graph components utilized for causal feature learning include

1. covariate Z refers to hidden variables that represent attribute properties;
2. outcome Y is the effect of treatment;
3. confounder W is a causal variable that impacts treatment and outcome; and
4. treatment T refers to an intervention that is deliberately applied to an attribute.

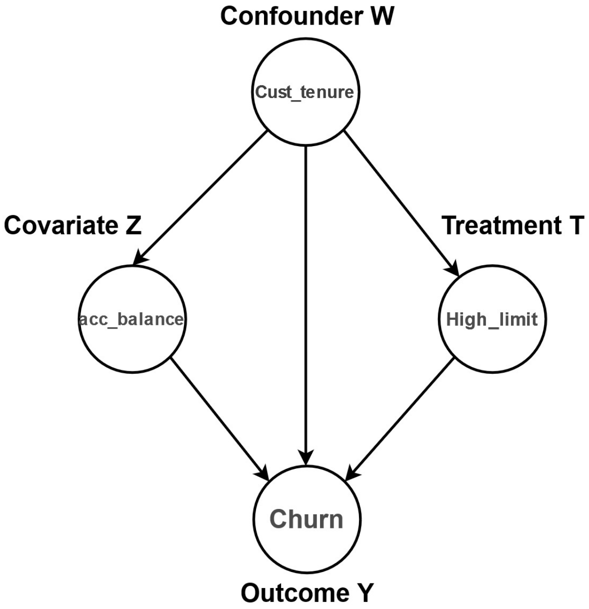


Figure 5.2: Example causal graph for churn

5.2.5 Causal Inference with DoWhy

The DoWhy Python library facilitates causal reasoning, allowing causal models to be developed with graphical representations. Even with partial graphs, DoWhy treats unmodeled variables as confounders, determines all methods to identify a causal effect, and employs graph-based rules and do-calculus for estimation. DoWhy is supported by back-door criteria and instrumental variables, and provides non-parametric confidence intervals and permutation tests for statistical validation. It also supports strategies such as adding control variables, using instrumental variables or implementing sensitivity analyses to address potential confounding and produce more reliable causal estimates. A key feature of DoWhy is its suite of refutation methods, which critically assess causal estimate validity.

5.2.6 High dimension feature space

Datasets commonly include more features than observations for each member, which can lead to overfitting the model. This is particularly common for financial data, where many features have low variance and correlation with the target variable.

Causal model performance depends on considering all causal features, possible covariates, cofounders, and attributes available in the data sources, and dropping low-importance features can increase bias in the model. In contrast, preserving all attributes leads to overfitting for predictive models. Hence, it is imperative to strike the appropriate balance between adverse and beneficial impacts from dimensionality [133, 134].

Many dimensionality reduction methods have been developed to address this issue, including feature dropping, wrapper methods, and feature importance with RF. This study employed recursive feature elimination (RFE) [135], obtaining robust results to overcome high-dimensional data while retaining possible influential cofounders and causal variables. The RFE algorithm is widely employed for ranking features since it provides a method to eliminate features with low weight and establish a threshold and a set number of top-ranked features.

RFE removes the feature with the smallest ranking criterion using the $DJ(i)$cost function,

$$DJ(i) = (1/2)\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \tag{5.2}$$

where $D\omega_i$ is an assigned weight. Evaluating feature weights involves evaluating the impact of altering the cost function output in weight $D\omega_i$ by eliminating a particular feature i using the iterative RFE procedure [135]. The ranking weight criterion for expanding J in Taylor series to second order can be expressed as

$$J = \sum_{x \in X} \|\mathbf{w} \cdot \mathbf{x} - y\|^2, \tag{5.3}$$

where J (5.3) is the cost function.

The features are divided into subsets F_m for each iteration, from the lowest to the highest ranked. This thesis also proposes an algorithm to combine SMOTE and RFE algorithms to circumvent minority class and high-dimension feature space issues. This strategy ensures that attributes with low variance and weaker predictive capabilities are removed from the collected data without significantly impacting the causal inference framework.

5.3 Methodology

5.3.1 Problem statement

Data mining can potentially extract valuable knowledge regarding pattern extraction from diverse sources, and various feature engineering tools can extract concealed patterns from vast datasets [136, 137]. Figure 5.3, shows 12 real-world datasets analyzed following the proposed approach to ensure the model's effectiveness.

This research addresses a binary classification challenge focused on the minority class. Each member is represented by a vector with n components or features. The data pattern P for each member exists in an n-dimensional feature space and is categorized into either minority or majority class (class 1 or 0, respectively). Therefore, I define a training set of vectors $\{x_1, x_2, x_3, \ldots, x_k, \ldots, x_n\}$ with corresponding class labels $\{y_1, y_2, y_3, \ldots, y_k, \ldots, y_n\}$, where $y_k \in \{0, 1\}$. A discriminant function, $D(x)$, is employed to distinguish vectors with n components or patterns. The decision boundary is split into regions where $D(x) > 0$ and $D(x) < 0$, enabling classification for each sample into churn (1) or non-churn (0) categories,

$$\begin{array}{lll} D(x) > 0 & \exists & x \in \text{class}(0) \Leftrightarrow \text{acc\_close}_{t_w} > \text{acc\_close}_{t_w} - 1 + t_w \\ D(x) < 0 & \exists & x \in \text{class}(1) \Leftrightarrow \text{acc\_close}_{t_w} \leq \text{acc\_close}_{t_w} - 1 + t_w \\ D(x) = 0 & \text{or acc\_close}_{t_w} = t_w & \qquad\qquad \text{decision\_boundary} \end{array} \tag{5.4}$$

where x are input patterns or vector components; t_w is a 6 month time-window; $acc\_close\_t_w$ and $acc\_close\_t_w - 1$ are the current and previous 6-month time windows, respectively. Hence, a linear discriminant function is constructed by calculating the sum of the training patterns and bias,

$$D(x) = w.x + b, \tag{5.5}$$

Where w is the weight for a pattern and b is bias. Since a singular criterion for defining the decision boundary has been established, a linear discrimination function can effectively distinguish the classes without introducing errors.

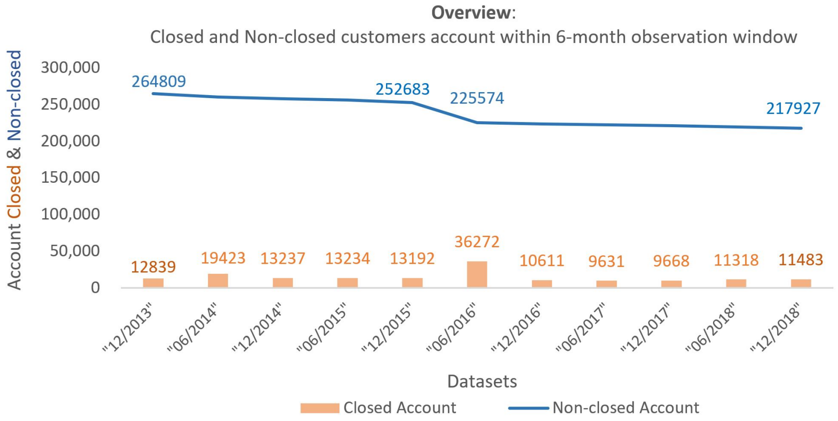


Figure 5.3: Member accounts (closed and non-closed) for 6 months

Equation (5.4) defines a member as a churner if they closed their account during the subsequent 6-month time window. Consequently, a binary classification is applied to each client indicating account closure (1), and 0 indicates its continuation within the following six-month period.

The data mining approach must align with the problem's structure, as defined by the need to analyze churn in member accounts that were either closed or remained open within a 6-month timeframe. two primary inclusion criteria were established to streamline the process and reduce dimensionality by eliminating excessive and noisy data. Only customers with account tenure longer than six months are retained, and those with account balances below $1500 are considered low-engagement members and are excluded. The observation window in Figure 5.4 showcases the features used to predict whether user will churn or not within the subsequent 6-month outcome window.

5.3.2 Churn predictive method

In assessing the efficacy of my novel DFF NNs algorithm, I compared Seven state-of-the-art classifiers were compared on datasets featuring 12-month observation period and 6-month forecast horizon, as detailed in the problem definition, to assessing the proposed novel DFF NNs algorithm efficacy. Figure 5.5 shows the operational flow for the proposed model. The algorithms were used for both training (80%) and evaluation (20%) (post feature extraction), extracting 193 features from a pool containing 124,363 instances.

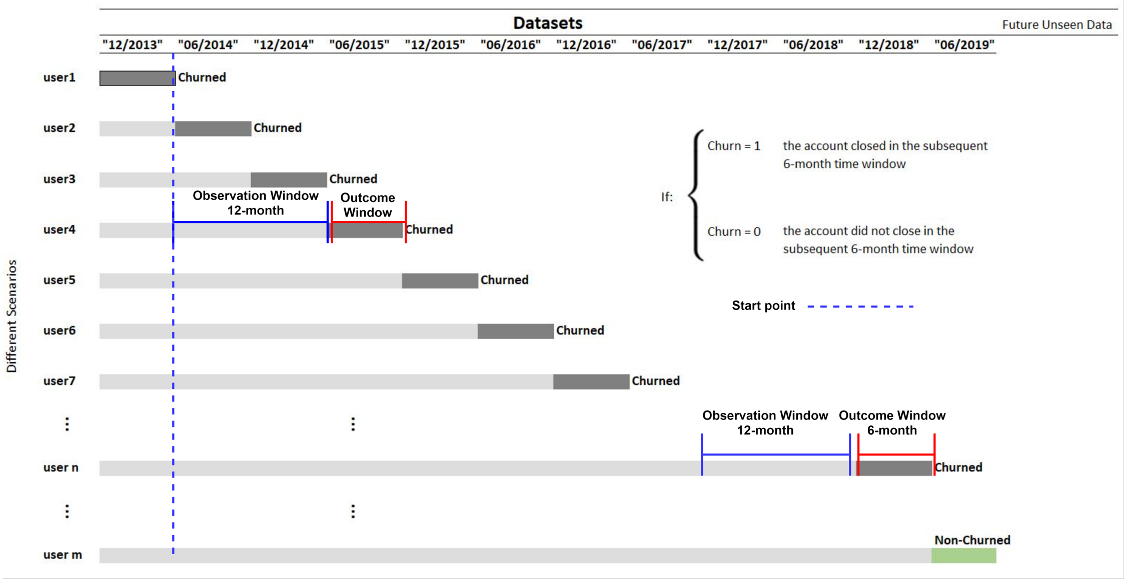


Figure 5.4: Proposed method to extract feature values using a sliding observation window onto a dataset. The observation and outcome time windows = 12 and 6 months, respectively.

The synthetic minority oversampling technique (SMOTE) [138] was applied during preprocessing to create synthetic samples for the minority class, balancing the training set with an equal number of data points (14031) in each class (1 or 0). Experimental results confirm that the model performance improved with SMOTE.

A majority voting ensemble mechanism was also applied from the Scikit-learn library [139] to boost classifier accuracy. Both ensemble hard and soft voting modalities were explored to assess supervised models. Hard voting tallies the votes from each classifier to determine the prevalent outcome, whereas soft voting assigns weights to predictions based on classifier significance, declaring the winner as the label with the greatest sum of weighted probabilities [139].

Table 5.2 shows the ensemble of ANNs employed to tackle the research question, deriving the fine-tuned ensemble ANN architecture by hyperparameter optimization.

5.3.3 Causal inference method

Causal graphs were employed to decode assumptions and identify dependency level granularity between features, using the DoWhy python package [127]. The causal model first scored attributes according to some measure of their importance for the predictive model to identify potential causal contributors, then feature weighting was implemented

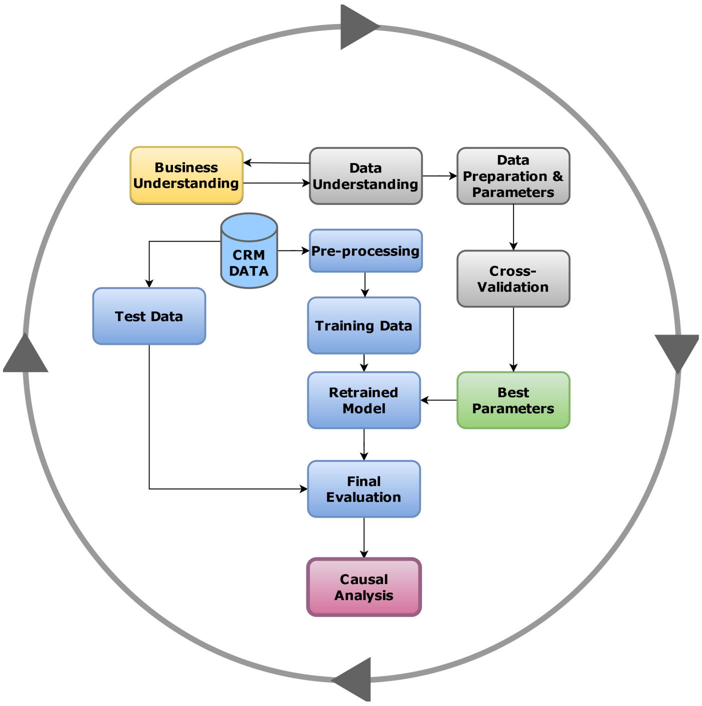


Figure 5.5: Framework overview for the proposed churn predictive model

Table 5.2: Ensemble ANNs network architecture for predicting churn.

Network type	Deep ANN-1	Deep ANN-2
Hidden layers	4	4
Dense activation 1,2,3 and 4	tanh, and 3$\sqrt{ó}$ Relu	tanh and 3$\sqrt{ó}$ Relu
Dropout 1, 2,3 and 4	0.2, 0.2, 0.2, and 0.2	0.4, 0.4, 0.4, and 0.4
Output activation function	Sigmoid	Sigmoid
Learning rate	0.000474718	0.000012
Epochs	100	100
Batch size	512	512
Optimization algorithm	ADAM	ADAM

based on predictive power between close attributes with different labels, SHAP value, and partial dependence plots (PDPs). Shapley values play a pivotal role by quantifying the contribution of each feature to the predictive model's output. By integrating Shapley values, the model assesses how the presence or absence of a specific feature affects the prediction outcome, thus isolating the impact of each attribute. Consequently, Shapley values provide a robust statistical basis to prioritize features for deeper causal analysis, ensuring that the causal model focuses on attributes that are most likely to drive changes in the outcome variable.

Figure 5.6 shows an example where the highest feature weights are the most powerful predictors. Feature influence on predictions and the correlation between key features and prediction accuracy were used to form preliminary causal hypotheses. These hypotheses were subsequently formulated in a directed acyclic graph (DAG) founded on prior knowledge from correlation assessments and influential predictor evaluations.

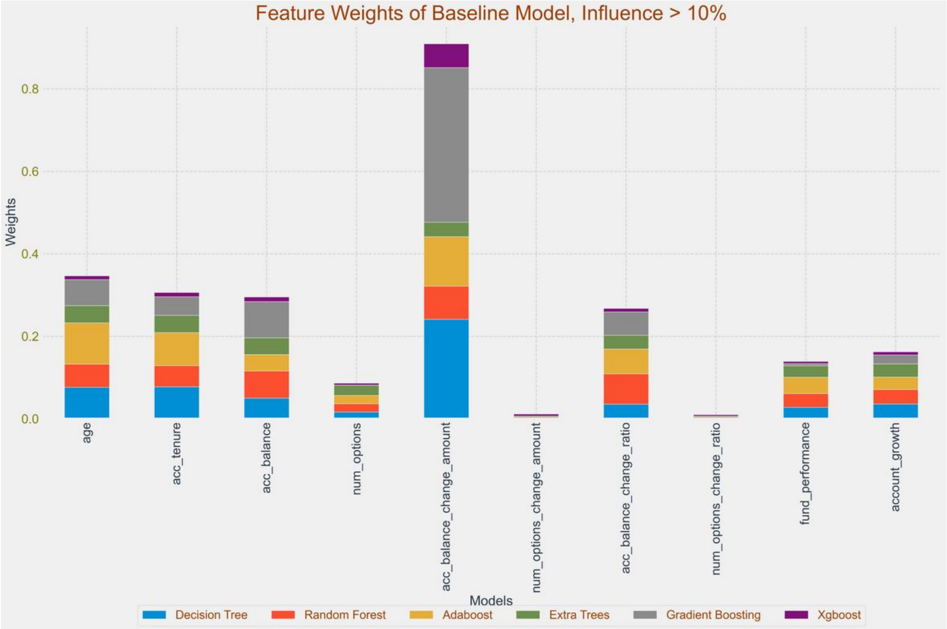


Figure 5.6: Example case for the 10 highest predictive features with influence > 10%

This thesis attempted to combine the ML and causal model to identify and incorporate influential predictors into the proposed causal inference framework, creating a causal inference framework with sturdy results built. A dependency architecture among predictor variables was also constructed based on relationships among independent variables. The following procedure was followed to model the causal inference.

1. Construct a causal graph by creating an underlying explicit causal graph for each causal assumption.

2. Identify causal effect by extracting target estimates based on influential predictors and causal graph observation.

3. Estimate causal effect employing a backdoor criterion, and subsequent permutation test to examine the estimated effect's statistical significance.

4. Refute the obtained estimate to validate the causal effect of the estimate.

5.4 Experiment Outcomes

5.4.1 Materials

Experiments were conducted using twelve datasets sourced from a local finance company containing information about customer accounts, demographics, customer engagement, and financial data. These datasets comprised approximately 250,000 examples, each containing 88 features, comprising 71 numerical and 17 nominal variables.

5.4.2 Churn Prediction Method

The churn prediction models employed in this research utilized advanced machine learning techniques to analyze and predict member behavior. The models were rigorously tested against various datasets to ensure robustness and accuracy. The results indicate that combining ensemble ANNs and feature selection techniques like RFE significantly improved the prediction accuracy. This subsection discusses the performance metrics, the effectiveness of the SMOTE technique in handling class imbalances, and how these methods contribute to identifying at-risk members effectively.

5.4.3 Causality Analysis Method

The causality analysis focused on identifying the root causes of churn to inform targeted intervention strategies. By employing advanced causal inference methods, including DoWhy for constructing and evaluating causal models, this research identified key factors influencing churn. The findings reveal that certain features such as account tenure and service usage patterns directly impact member retention. This subsection discusses the implications of these causal relationships and how they can be leveraged to formulate effective churn mitigation strategies.

5.5 Results and Discussion

This section presents the findings from the analysis of churn prediction methods and causality analysis methods. Each subsection elaborates on the respective methodologies applied and discusses the implications of the results in the context of enhancing member engagement and reducing churn.

5.5.1 Churn prediction results

The results of my experiment are presented in Figure 5.7 shows the experimental outcomes for the most recent data and outcome windows. RF outperformed the other algorithms, achieving AUC = 80%. The proposed ensemble ANNs and RF achieved similar performance, with 7.5% improved AUC compared with logistic regression. Ensemble ANNs achieved maximum prediction accuracy on test data and were comparable with current best-practice classifiers. The evidence base for exploiting ANNs reduced time-consuming feature engineering, requiring expert knowledge, for these specific financial datasets. Table 5.3, shows the low Cohen kappa score = 0.86 for the proposed model, confirming no significant difference between the null error and test accuracy outcomes. The proposed algorithm also achieved Matthews correlation coefficient (MCC) range < 0.45, confirming the model's reliability.

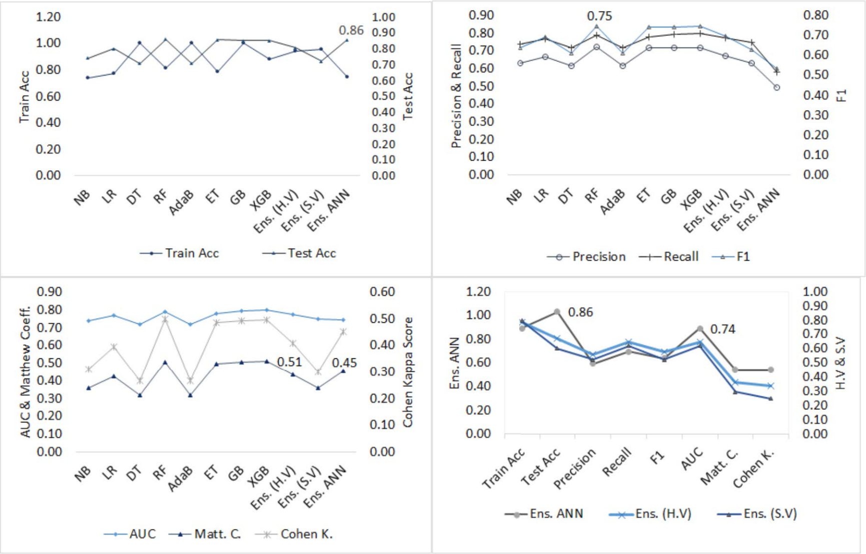


Figure 5.7: Churn predictive model performance using several evaluation metrics and best-practice classifiers. The highest test accuracy was achieved using the proposed ensemble ANN.

Descriptive methods in statistical analysis define the predictive power for features and their contribution to improving model output from its base (average output for the training dataset) to more meaningful values. Figure 5.8 uses SHAP plots to extract the feature prediction impact on model performance, where red features increase and blue features reduce the base value model outcome, e.g. feature acc_balance_change_amount impact reduces, and feature sg_recency increases, model predictive power. Figure 5.9 uses the PDP plot to show that the most potent predictors all have significant impact on model

Table 5.3: Model performance metrics for ten classifiers

Model	Test accuracy	Cohen kappa	Matthew coeff.
Naive Bayes	0.74	0.36	0.31
Logistic regression	0.80	0.43	0.40
Decision tree	0.71	0.32	0.26
Random forest	0.86	0.51	0.50
AdaBoost	0.71	0.32	0.27
Extra trees	0.86	0.49	0.49
Gradient boosting	0.85	0.51	0.49
XGboost	0.85	0.51	0.50
Stack ensemble (H.V)	0.81	0.44	0.41
Stack ensemble (S.V)	0.72	0.36	0.30
Ensemble ANNs (ours)	**0.86**	0.45	0.45

performance. Thus, the PDP and SHAP plots confirm that acc_balance_change_ratio, login_recency, acc_tenure, cust_tenure, and account_growth_change constitute be strong causal estimators under the causal hypothesis.

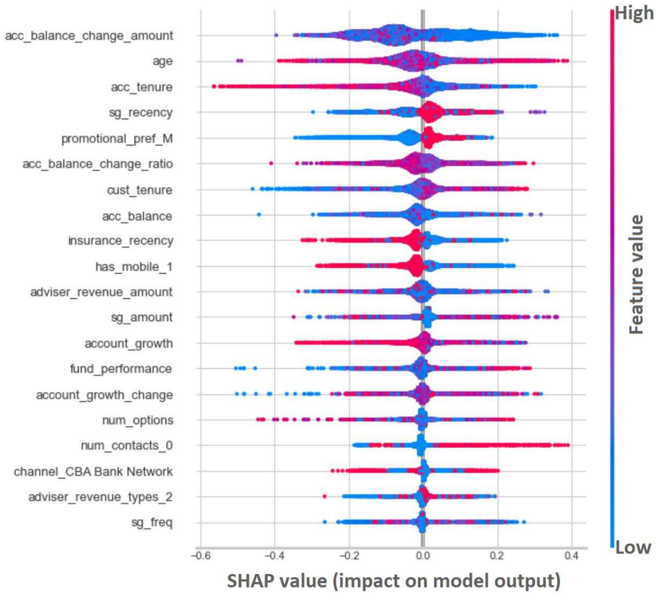


Figure 5.8: The SHAP graph illustrates feature impacts on prediction outcomes. Feature acc_balance_change_amount and sg_recency are the most important features that improve prediction accuracy.

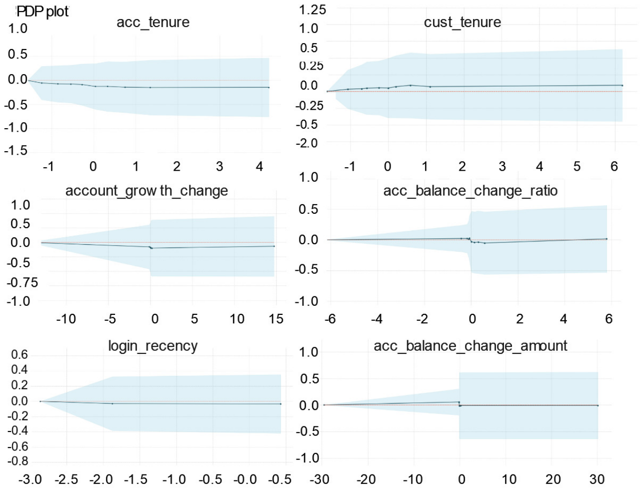


Figure 5.9: The PDP plot shows that six causal assumptions can be related to features since they are identified as the highest predictive power variables for model prediction outcomes.

5.5.2 Causality analysis results

Treatment causal effects on churn outcome were identified based on the initial assumptions, keeping other potential effects constant while changing the target treatment. For example, linear regression estimation achieved that estimated effect = -0.033853 corresponds to churn probability reducing $\approx$ 3% when the member has lower account growth rate. To verify this assumption, i.e., that it is true, the new estimation effect should not alter significantly. Therefore, the data subset refuter was applied to disprove these estimates by rerunning them on a random subset of the original dataset. The refuting method outcome = -0.033920, almost identical to the estimation result. Thus, the assumption was correct that high account tenure was a causal feature for churn outcome.

Treatment effects on the outcome depend on value changes for the treatment variable. The effect impact degree is determined by statistical analyses, and there are many suitable statistical methods for assessing the causal effect. The causal experiment employed propensity score-based inverse weighting (PSIW) and PSM methods [128]. Figure 5.10 displays the causal relationships between various attributes and their impacts on customer churn, confirming the interconnected factors influencing churn decisions. The graph validates the causal assumptions, showing how specific variables directly affect churn outcomes. Table 5.4 shows final and churn probability estimations. Mean estimation $\sim$ 0.15 for variable sg_recency, which is equivalent to increasing the churn probability ∼ 15% when the customer has longer time-period since the last day of super guarantee (SG) contribution. Similarly, churn probability increases ∼ 3% when the customer has negative account growth rate, mean estimation ∼ 0.03 for account_growth. Although the causal analysis outcomes support the assumptions regarding confounding factor identifications are accurate to a high degree, several limitations remain in terms of analyzing the identified confounding effects with other popular causal inference methods, such as counterfactual analysis. Counterfactual causal analyses offer a distinct approach to causal inference that can provide valuable insights beyond propensity score-based methods such as PSIW and PSM. These different approaches complement each other and can be valuable for gaining a comprehensive understanding of causal relationships in complex datasets.

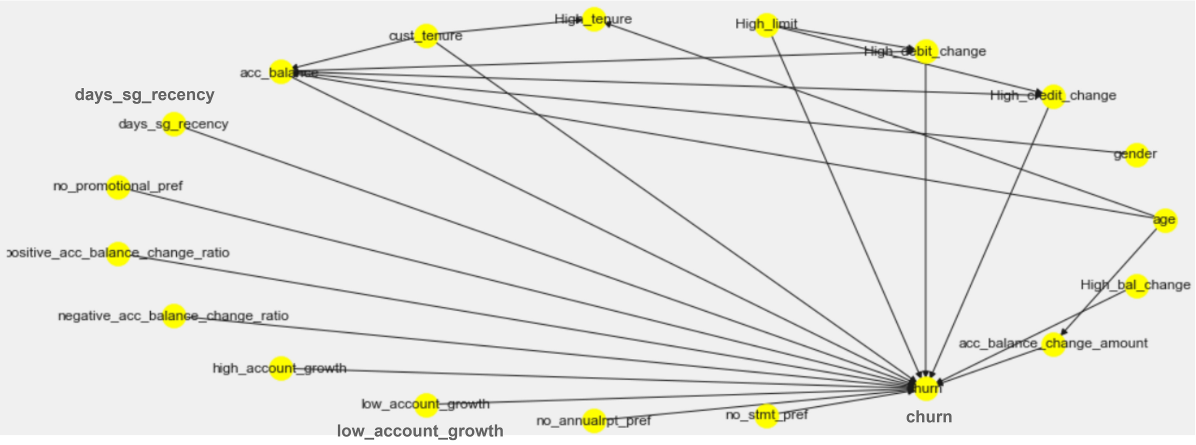


Figure 5.10: The causal graph illustrates the causal link between attributes and causal assumptions, confirming the interconnected factors influencing churn decisions.

Table 5.4: Causality analysis illustrates assumptions that have causal effects on customer churn are valid

Causal variable	Estimate effect	Data subset refuter	Churn probability
high_bal_change	−0.123401	−0.122474	reduced by ∼12%
high_acc_balance	−0.091698	−0.091612	reduced by ∼9%
low_account_growth	−0.033853	−0.033920	increased by ∼3%
annualrpt_pref	0.144440	0.144457	increased by ∼14%
stmt_pref	−0.142732	−0.142614	decreased by ∼14%
high_cust_tenure	−0.027969	−0.081893	decreased by ∼3%
high_sg_recency	0.156396	0.156396	increased by ∼15%
promotional_pref	−0.086401	−0.088061	decreased by ∼8%

5.5.3 Discussion

Customer loss or churn is a common challenge faced by member-centric organizations. However, this thesis has shown that it is possible to mitigate churn by identifying and investing in customers who are at risk of leaving, ultimately maintaining acceptable retention levels. One of the core contributions is the development of a novel and comprehensive churn propensity model. This model incorporates various features and leverages advanced techniques, including SMOTE sampling, RFE, ensemble artificial neural networks, and causal inference methods via DoWhy. Combining these approaches enables accurate churn predictions, and helps understand the root causes for churn. Experimental outcomes confirmed the proposed ensemble ANNs model effectiveness, and the proposed ensemble ANNs achieved the highest accuracy on test data compared with ten current best-practice classifiers. This emphasizes the importance of considering advanced machine learning techniques for churn prediction, as they can significantly outperform traditional methods.

The proposed causal inference model provided valuable insights into factors contributing to customer churn. Variables such as recent SG_contribution, changes_annual_report, statement_preferences, account_growth_rate, and balance_amount were identified as confounding factors with high degree of belief, i.e., these variables play exert significant influence on churn outcomes. For example, customers with active accounts for over a year exhibited ~ 3% reduced churn rate, aligning with expert knowledge; and high account balance greater than 100,000 AUD was associated with ~ 9% reduced churn probability.

Future work will refine and expand the proposed framework. One avenue for improvement is to extract patterns with smaller outcome windows, which could lead to more efficient prediction results. Another aspect will be to explore applying counterfactual causal analysis in churn prediction, which could provide useful alternative perspectives on causal relationships within the context of churn, ultimately yielding in-depth insights into customer behavior and retention strategies.

This research advanced our understanding of churn prediction and causal inference in member-centric organizations and developed a robust framework for addressing customer churn by combining data mining techniques and causal inference analysis. Insights gained from this study will help inform proactive retention strategies and ultimately help organizations reduce churn and enhance customer satisfaction.

The Python implementation for the proposed framework, causal analysis result, and visualization are available on the GitHub repository[1] in the footnote.

[1] https://github.com/DavidHason/Causal Analysis

5.6 Summary

Chapter 5 investigated causal analysis for member churn by integrating ANNs with a causal inference model. The core objective was to explore causal factors behind member churn and gain more in-depth understanding of this phenomenon. This approach first emphasizes the connection between member engagement and churn, highlighting the need to understand how to retain individual users effectively. The methodology covers data preprocessing, feature selection, and machine learning algorithms, focusing on addressing class imbalances and a causal mode using propensity score-based methods, such as PSIW and PSM. Empirical results confirmed the proposed ensemble ANNs model superior performance for churn prediction, and key causal variables, confounders, and covariates, such as account_growth_rate and balance_amount were identified. The discussion highlights future research possibilities, including applying counterfactual causal analysis to gain more profound insights into churn dynamics.

Thus, Chapter 5 provides a significant contribution to understanding member churn and valuable insights for improving member engagement and retention.

CHAPTER 6

Churn Prediction via Multimodal Fusion Learning: Integrating Member Financial Literacy, Voice, and Behavioral Data

This chapter proposes an innovative multimodal fusion learning framework to determine churn risk levels in financial service organizations, integrating customer sentiment, financial literacy (FL), and behavioral data to produce more accurate and unbiased churn predictions.

6.1 Background and Motivation

A thorough grasp of member engagement and attrition is crucial for financial organizations [2, 3]. In-depth understanding of member behaviors and needs will require moving from traditional methods to advanced analytical frameworks. Recognizing the psychological elements that influence customer decisions also vital for financial organizations. Cognitive biases, e.g. anchoring bias, availability heuristic, and bandwagon effect, can significantly impact customer perceptions and decision-making regarding products and services [4]. Misguided interactions or poor presentation of complex financial products can amplify these biases, steering customers to focus on potential risks rather than benefits. Therefore, it is crucial for financial service providers to understand these biases and to tailor their product design, marketing, and CRM strategies effectively, reducing negative biases and churn rates.

One robust solution that helps mitigate cognitive biases and elevate the extent of member engagement in financial organizations is utilizing customer voice (CV), financial literacy (FL), and CRM data. Member engagement and churn are two sides of the same coin. Therefore, this methodological approach aids in formulating effective strategies for member retention and promoting enduring loyalty for continuous organizational growth. Members with robust financial understanding can make more informed decisions with less cognitive bias, reducing the churn likelihood due to misunderstandings or unfulfilled expectations.

On the other hand, customer data is undergoing sustained and exponential growth, and traditional churn analysis methods, which heavily rely on historical transactional and demographic data, are no longer sufficient. More comprehensive approaches are urgently required to address this data surge across various modalities.

Previous studies have established many useful insights into member behaviors regarding churn. Nonetheless, these studies have typically missed the opportunity to connect member engagement with churn by considering a comprehensive perspective that covers interactions, emotions, FL, and CRM data. Past methods often relied on singular data sources, such as transactional, demographic, and textual data, which do not offer a well-rounded view of member behavior. Ongoing research faces three primary obstacles.

1. Transactional data used in many previous and recent studies only mirrors predictive results without investigating the root causes for churn.

2. Demographic data, although widespread and easily available, is static and fails to reflect customer satisfaction evolution.

3. Social media data, which is limited to textual content, provides a noisy and impersonal dataset that lacks insight depth, such as those from voice interactions and financial behavior insights.

Therefore, this chapter proposes a multimodal modeling framework that captures the complex layers of member engagement, providing a foundation for developing a multimodal hybrid fusion learning model that not only combines FL metrics and behavioral indicators with emotional nuances of voice data, but also integrates the critical element of member engagement. This integration markedly improves reliability, accuracy, and model bias, and provides a cutting-edge model employing diverse neural network designs, enhanced data enrichment methods, and sophisticated emotion recognition algorithms.

The proposed approach employs the SMOGN-COREG supervised algorithm to measure FL from customer financial information, and the baseline churn model utilizes an ensemble ANN, complemented by oversampling methods, to forecast churn in vast financial datasets. Additional aspects include a proposed SER model based around VGG-optiVMD, a novel speech signal processing algorithm, to analyze customer emotions through acoustic features such as pitch, energy, and tone.

This thesis introduces a groundbreaking approach to churn prediction, incorporating multimodal machine learning techniques for comprehensive understanding of customer churn. The proposed method considers multiple aspects of the customer experience, from service interactions to product engagement. Multimodality in machine learning integrates different input types recorded in various media formats into a single model, where these inputs are not straightforwardly interchangeable using an algorithm [6].

Thus, thus thesis proposes a multidimensional fusion learning framework to collect these inputs into a coordinated feature representation space and subsequently applying decision-level fusion to assess churn risk.

The approach for sentiment analysis differentiates between paralinguistic methods using customer voice (CV) analysis and linguistic/text-based methods. Although natural language processing (NLP) can transcribe and analyze customer service calls, it often misses meta-information conveyed through tone, pitch, and loudness, aspects that non-linguistic emotion detection using speech signal processing can capture more robustly. Negative emotions from customers can be strong indicators for product dissatisfaction, placing them at higher churn risk, and meriting particular attention.

Financial literacy is an essential factor in churn prediction within the financial services industry. Customers with higher FL are more likely to understand and choose suitable financial products, reducing dissatisfaction and churn likelihood. On the other hand, those with lower FL may struggle with these choices, leading to dissatisfaction and possible defection to competitors with seemingly superior offerings.

6.1.1 Key contributions

Previously proposed unimodal models' effectiveness was assessed and several significant contributions introduced.

- This thesis pioneers a multimodal hybrid fusion model that fuses distinct CRM, CV, and FL databases. Embedding an essential layer to churn prediction by detecting customer emotion from vocal attributes and assessing financial competence from account performance and survey data.

- A specific coordinated feature representation space and translation matrix to address output value heterogeneity across modalities using predefined logical propositions, with the potential to incorporate future modalities, such as textual features.

- This thesis empirically proves substantial correlations between negative emotions, low FL, and increased churn risk.

6.2 Preliminary Knowledge

Recent scientific and technological advances have led to a new where data are abundant and accessible in numerous formats. Consequently, multimodal learning has become increasingly popular and significant field within the deep learning domain. The fundamental principle of multimodal learning is to process and interpret a diverse range of information types, which is crucial for a comprehensive understanding of real-world objects and phenomena.

A singular representation mode, whether it be visual, auditory, or textual, provides a limited perspective, lacking the broad representation required for comprehensive understanding. Therefore, multimodal fusion learning has emerged as an important innovation, aiming to consolidate diverse data streams into a unified analytical framework. This integration enhances the depth and breadth of data analysis and brings more human-like dimensions to problem solving by simulating how humans assimilate information from various senses.

One of the primary challenges for multimodal learning is to efficiently fusion features from different modalities, while preserving the unique characteristics for each modality to minimize information loss. This section describes multimodal analysis evolution, and briefly discusses the main approaches to multimodal fusion, prevalent models, and their specific applications in the current technological landscape.

The term "multimodality" has its roots in philosophy and the arts, dating back to the fourth century BC, where it was used to define expressive and rhetorical strategies that combined different content forms [140]. Fast-forward to the twentieth century and the growth of the internet and mobile technology has positioned multimodal data as the most frequent form.

Multimodal fusion has gained considerable traction with the rapid progression of deep learning. It can be used in many areas and generally helps to improve application efficacy. Notable applications include face recognition [141], where multimodal learning facilitates the analysis of facial features by integrating visual data with other relevant modalities; visual question answering [142], combining textual and visual data to understand and respond to queries related to images.; image captioning [143], where multimodal learning helps to generate descriptive text for images, leveraging both visual cues and linguistic patterns; sentiment analysis has been enriched by the ability to analyze and interpret sentiments from text, audio, and video data collectively [144].; and finally, multimodal retrieval offers far more efficient searching and retrieving information becomes by indexing and processing multiple data forms concurrently [145]. It is anticipated that multimodal fusion learning will continue to revolutionize the way we interact with technology and extract meaningful insights from the ever-growing expanse of data.

6.2.1 Definition and evolution of multimodal learning

"Modality" refers to the various ways humans perceive the world, such as sight and sound. In computing, a modality represents any data type, like sounds, images, or text. Unimodal learning translates this data into numerical vectors for computational analysis, whereas multimodal learning enhances the comprehension of such diverse data by capitalizing on their complementary attributes and reducing redundancy, often combining images, text, and audio for a more integrated approach to learning.

Multimodal learning approaches evolution has been significant in numerous intelligent data processing areas since the 80s. McGurk et al. clarified visual element influences on speech perception in 1976, setting the foundation for audio-visual speech recognition [146]. This discovery, known as the McGurk effect, inspired many computer scientists to explore multimodal speech recognition systems to integrate visual and auditory information, such as lip-sound speech recognition systems [147], notably enhancing accuracy over audio-only systems. Atrey et al. [148] categorized multimodal fusion techniques and their level in 2010. Wang [149] subsequently introduced deep multimodal hashing with orthogonal regularization constraints, aiming to minimize information redundancy within multimodal representations. Zhang et al. [150] and Wang et al. [151] greatly help develop cross-modal information matching and retrieval, Liu et al. [152] integrated visual and haptic data, which has been incorporated into the integrated robotic perception domain; and Fu et al. [153] advanced the field of semantic image annotation.

6.2.2 Multimodal fusion learning methods types

Fusion methods not tied to a specific model are categorized based on the timing of the fusion process: early fusion (EF) merges features immediately post-extraction, late fusion (LF) combines outputs from each model, and hybrid fusion blends the benefits of both approaches.

6.2.2.1 Early fusion

I have added two paragraph regarding why EF not always suitable method in this particular multimodal approach. To sum up, EF might not always be optimal for combining multimodal data due to its tendency to create redundant input vectors, leading to overfitting and increased computational complexity. Although neural networks can mitigate these issues with advanced techniques like dropout and regularization, EF can still dilute the unique characteristics of each modality, limiting the model‚Äôs flexibility and effectiveness in capturing complex cross-modal interactions. This underscores the need for alternative fusion strategies in complex applications.

6.2.2.2 Early fusion

Early fusion refers to the merging feature and data levels directly following feature extraction, often through a straightforward join operation on the feature sets. Figure 6.1(a) shows the framework for early fusion methods, where extracted features are immediately fused, followed by integrating features from different modalities for model training. This method is particularly beneficial when modalities are closely related, allowing exploitation of correlations and interactions between low-level elements of each modality. However, Hinton et al. [154] showed that extracting correlations at the feature and data levels presents a significant challenge, and in some instances, data from diverse modalities may only exhibit significant correlations at a more abstract level(s). Martinez et al. [155] argued that early fusion might not effectively exploit the complementary nature of different modalities, potentially leading to unnecessarily redundant input vectors, which could lead to issues such as overfitting or increased computational complexity. However, this concern might be less significant when using neural networks because they can learn to ignore redundant information through their inherent ability to prioritize relevant features during the training process. Moreover, advanced neural networks often incorporate techniques such as dropout, L1/L2 regularization, and batch normalization, which help mitigate the risk of overfitting even when faced with high-dimensional data.

Despite neural networks' ability to handle redundancy through feature prioritization during training, incorporating techniques like dropout and regularization to combat overfitting, Early Fusion (EF) may not always be the most effective strategy for combining multimodal data. EF merges all modalities at the onset, potentially diluting the unique characteristics of each data type and making the model inflexible, as all data must be simultaneously available in the same format. This can be particularly limiting when distinct modalities could benefit from specialized processing methods such as CNNs for images or RNNs for sequential data. Moreover, EF may hinder the network's capacity to learn more abstract and complex cross-modal interactions, which could be more effectively captured through separate or sequential processing. Therefore, while EF offers a simplified integration process and utilizes neural networks to manage data redundancy, it lacks the flexibility, specificity, and efficiency required for optimal learning from multimodal data in complex applications like churn prediction, highlighting the importance of exploring alternative fusion strategies like late or hybrid fusion.

6.2.2.3 Late fusion

Late fusion, also known as decision level fusion, involves independently training a distinct model for each modality before merging their outputs. This fusion category employs various methods, including Bayesian rule fusion, maximum fusion, mean fusion, and other rule-based approaches, to integrate outputs from different models [156].

This fusion method offers advantages over early fusion techniques by accommodating data synchronization and providing the flexibility to choose the most appropriate analysis method for each modality, such as using hidden Markov models for audio data and SVM for visual data. However, LF tends to overlook low-level interactions between modalities, which can complicate the fusion process. Figure 6.1(b) shows a fairly typical framework for LF methods, where the initial stage trains models on data from each modality separately, and the subsequent stage combines these models' outputs through a decision making rule.

6.2.2.4 Hybrid fusion

Hybrid Fusion methodologies combine the advantages of early and LF, but at the expense of increased model complexity and a more demanding training process. hybrid fusion is particularly well-suited for deep learning models due to their adaptable and varied structures, and has been extensively applied visual question answering and multimedia fields. For example, Ni et al. [157] proposed a hybrid fusion technique for multimedia analysis,

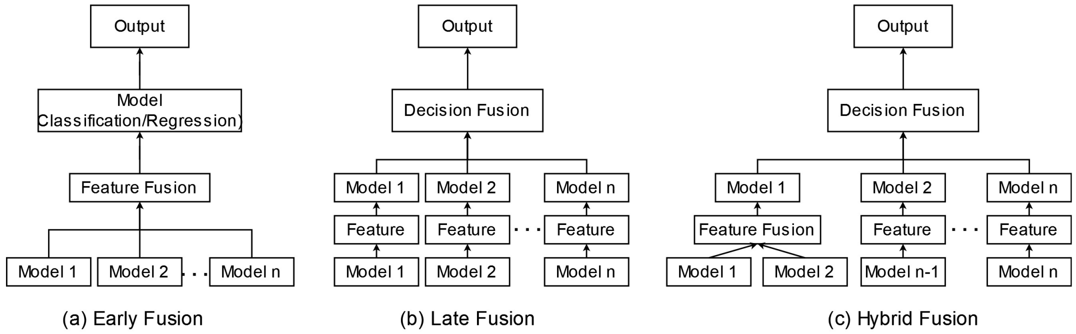


Figure 6.1: Multimodal fusion learning method types

introducing an image fusion methodology utilizing multiple back propagation networks. The proposed method merged video and audio signal components into an audiovisual deep neural network model to derive predictions, with outcomes produced by combining each model's predictions [158]. hybrid fusion strategy effectiveness significantly depends on the logical coherence of the combination approach, which is crucial for enhancing model performance. Figure 6.1(c) shows an example hybrid fusion method, integration early and LF strategies. Each fusion method has its strengths and limitations, early fusion is adept at capturing inter-feature relationships but may lead to overfitting; LF addresses overfitting but does not allow simultaneous training on all data; and hybrid fusion combines early and LF benefits, necessitating careful selection of appropriate fusion methods that aligns with the specific challenges of practical applications. This thesis proposed a hybrid fusion model and compared that model's effectiveness with late and early fusion methods.

6.2.2.5 Feature representation and challenges

Feature and representation terms are used synonymously, denoting a vector or tensor that encapsulates input data, whether an image, audio clip, word, or sentence [72]. A multimodal representation integrates data from diverse sources, and representing multiple modalities can often present challenges, including merging data from diverse origins, dealing with varying noise levels, and addressing data omission. Constructing meaningful data representations is essential for multimodal applications and forms the basis for any model.

Therefore, maintaining semantic similarity presents a significant challenge in multimodal learning due to the complexity to recognize sophisticated relationships among various modal inputs. This heterogeneity is vital, since ANNs aim to fuse attributes from different modes into a unified feature representation space. For example, language is often represented as symbols, while sounds and pictures are shown as signals. Representation quality significantly impacts machine learning model efficacy, e.g. advances in speech recognition and visual object categorization [159, 160]. Several key characteristics of superior representations, including smoothness, spatial and time coherence, sparsity, and natural grouping, have been highlighted in previous studies [161]. The feature representation space should reflect conceptual similarities, be readily derived when certain modes are absent, and allow for completing missing modes based on what's seen [162].

For example, the audio signal processing employed in this thesis has moved away from traditional acoustic features, such as Mel-frequency cepstral coefficients, and more towards data-driven deep neural networks for speech recognition and recurrent neural networks for analyzing paralinguistic features [88, 159]. Despite extensive unimodal representation, where most previous multimodal representations were simply concatenated unimodal models, the landscape is currently evolving rapidly [163].

6.3 Methodology

The proposed method analyzes customer behavior through three distinct modalities: baseline customer churn, SER, and FL. Subsequent sections provide detailed discussions of each modality.

6.3.1 Data prepocessing

Several distinct real-world data sources were employed to assess the proposed model's effectiveness, including historical financial data, CRM demographic information, and customer voice (CV) recordings, as shown in Table 6.1. A de-identification technique was employed to ensure privacy while retaining emotional nuances, and the CV samples were substituted with analogous samples from established emotion databases. The constructed CV database exploited correlations between negative emotions and high churn risk, financially illiterate customers and the converse for positive emotions with low churn risk customers. Emotion classification was extracted using the EMODB database, a recognized benchmark for emotion categorization from voice data [114], and the FL database combined data from financial transactions and customer surveys, resulting in an enriched training dataset with (4154, 140) dimensionality.

Table 6.1: Databases employed

Sub Database	**Specification**		
	No. datasets	***Sources***	***Attributes / Size***
DB1	1	Financial networks	68 / 64K
DB2	12	CRM	87 / 294m
DB3	1 [a]	Audio	4 / 110 min. voice

[a] EMODB database

To ensure meaningful multimodal learning and alignment of individual customer records across the diverse datasets, we employ a logical mapping method that systematically associates customer CRM data, Financial Literacy (FL), and Customer Voice (CV) data. This approach is critical to mitigating bias in multimodal modeling by accurately matching related modal inputs for each customer. For instance, the integration of these datasets follows predefined criteria that correlate negative emotions or complaints in the CV records with corresponding CRM and FL data. Specifically, the alignment is conducted under rigorous conditions considering multiple financial and engagement metrics such as account balance changes amount, account tenure, current account balance, engagement factors, and financial literacy scores. These metrics reflect crucial aspects of customer interaction, financial health, and overall member engagement.

For a customer exhibiting negative emotions like anger or disgust, our model filters and aligns CRM entries where the account balance change is significantly lower than the mean, tenure is less than six months, engagement factors are notably below average, and the financial literacy score is below 25 out of 100. This meticulous alignment ensures that the sentiment expressed in CV data directly corresponds to the customer's actual financial behavior and interaction history as recorded in CRM and FL datasets. Such detailed and context-sensitive alignment prevents the juxtaposition of incongruent data,Äîsuch as pairing a CRM entry of a satisfied customer with a CV record of a customer complaint,Äîthereby enhancing the accuracy and effectiveness of the multimodal learning process.

6.3.2 Modalities

6.3.2.1 Customer financial literacy modeling

Considering the significant role that inadequate FL plays in customer attrition, the first modality focuses on FL. The SMOGN-COREG semi-supervised regression model was developed previously (see chapter 3) to quantify FL from large and unlabeled real data sets collected from financial network data. The learning process was enhanced by creating synthetic samples for minority classes. The SMOGN technique was integrated with COREG, a non-parametric multi-learner semi-supervised regression, to improve model performance by evening out the response variate distribution, i.e., amplifying the presence of infrequent yet critical instances within the data [118]. The output of this model is an FL score ranging from 0 to 1.

6.3.2.2 Emotion recognition modeling

The second modality employs voice-based sentiment to distinguish between positive and negative member emotions during call center engagements. The proposed model analyzes vocal attributes, including tone, pitch, and rhythm, extracting acoustic features from the percussive and harmonic components of the Mel spectrogram [82]. A CNN-VGG16 model pretrained to construct a framework for recognizing emotions from a CV was incorporated with a feature map generator function in the proposed framework to extract harmonic and percussive components by applying a median filter to the signal spectrum axes. The log Mel spectrogram was computed after averaging these components and creating two feature vectors, and the subsequent 2D image feature map was used as input to the CNN-VGG16 network to classify emotions into binary outcomes, i.e., positive or negative.

6.3.2.3 Baseline churn modeling

The third modality assessed customer churn risk by analyzing demographic and financial behavior data extracted from CRM systems. This model combined recursive feature elimination, SMOTE, and deep learning ANNs to develop a classifier for predicting churn. The proposed model utilizes historical data spanning the previous 12 months to forecast churn likelihood for the following six-month period. In this framework, a ’churner’ is identified if an account is discontinued within the six-month forecast window, with binary outcome (0 or 1) signifying an active or terminated account, respectively, for that upcoming period.

6.3.3 Proposed multimodal fusion learning

The proposed multimodal fusion learning model is adept at analyzing organizational member engagement through an extensive range of data sources, which helps to under-

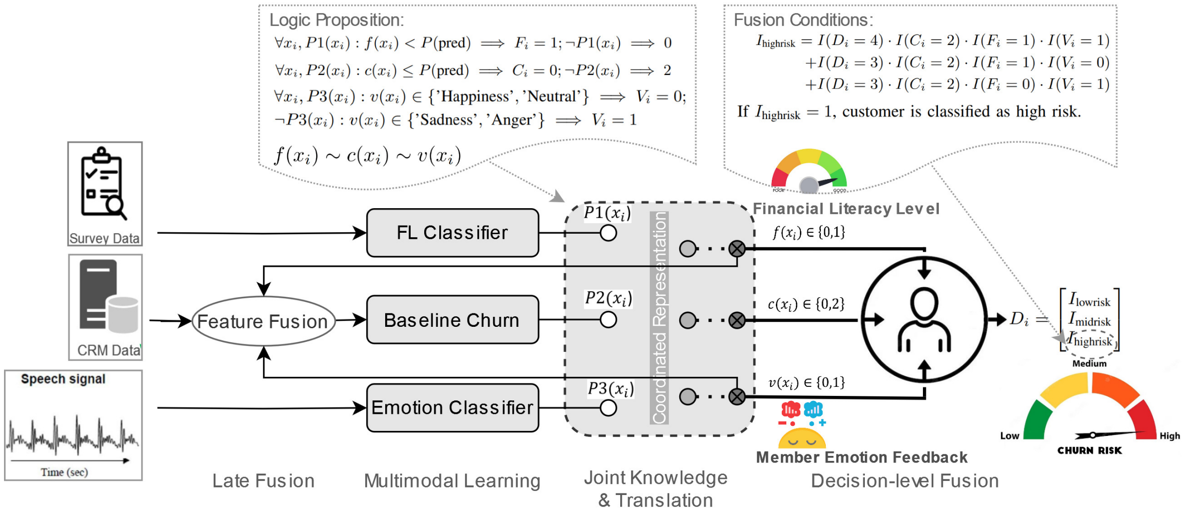

Logic Proposition:
$\forall x_i, P1(x_i) : f(x_i) < P(\text{pred}) \implies F_i = 1; \neg P1(x_i) \implies 0$
$\forall x_i, P2(x_i) : c(x_i) \leq P(\text{pred}) \implies C_i = 0; \neg P2(x_i) \implies 2$
$\forall x_i, P3(x_i) : v(x_i) \in \{\text{'Happiness'}, \text{'Neutral'}\} \implies V_i = 0;$
$\neg P3(x_i) : v(x_i) \in \{\text{'Sadness'}, \text{'Anger'}\} \implies V_i = 1$
$f(x_i) \sim c(x_i) \sim v(x_i)$
Fusion Conditions:
$I_{\text{highrisk}} = I(D_i = 4) \cdot I(C_i = 2) \cdot I(F_i = 1) \cdot I(V_i = 1)$
$+I(D_i = 3) \cdot I(C_i = 2) \cdot I(F_i = 1) \cdot I(V_i = 0)$
$+I(D_i = 3) \cdot I(C_i = 2) \cdot I(F_i = 0) \cdot I(V_i = 1)$
If $I_{\text{highrisk}} = 1$, customer is classified as high risk.
POOR
GOOD
Survey Data
CRM Data
Speech signal
Time (sec)
FL Classifier
Feature Fusion
Baseline Churn
Emotion Classifier
$P1(x_i)$
$P2(x_i)$
$P3(x_i)$
Coordinated Representation
Financial Literacy Level
$f(x_i) \in \{0,1\}$
$c(x_i) \in \{0,2\}$
$v(x_i) \in \{0,1\}$
$D_i = \begin{bmatrix} I_{\text{lowrisk}} \\ I_{\text{midrisk}} \\ I_{\text{highrisk}} \end{bmatrix}$
Member Emotion Feedback
Medium
Low
High
CHURN RISK
Late Fusion
Multimodal Learning
Joint Knowledge & Translation
Decision-level Fusion

Figure 6.2: Proposed multimodal hybrid fusion learning method workflow to integrate various modalities

stand the intricate interplay among various aspects of customer behavior. This framework comprises several elements: three independent unimodal models, a feature representation space, a mechanism for translating or mapping these unimodal outputs, and a hybrid fusion strategy. Figure 6.2 shows how the framework containing these three models predicts member FL, emotional response, and churn likelihood to categorize members into low (loyal or non-churner), mid (possible churner), and high (likely churner) risk. Broadly, a high-risk churn customer has limited FL, negative emotions towards company services, and demonstrates significant churn propensity.

The proposed model integrates a multi-level, hybrid fusion strategy incorporating early and LF principles. This hybrid fusion provides informative and enriched features for the base churn model by fusing predictive insights resulting from FL and SER unimodal models into the CRM dataset. A prototype multimodal learning model employing only LF was also developed to provide a benchmark to measure the proposed novel hybrid fusion strategy effectiveness against. Several late fusion multimodal methods were also employed for model comparison purposes by combining different unimodal models.

6.3.4 Feature representation space and translation

Identifying particular features unique to each modality is a complex task, given the natural variations in scale, measurement, and distribution. Therefore, a multimodal model was designed that merges these diverse features into a coordinated feature representation space while preserving their distinct contributions to churn prediction. It can encapsulate data similarities, capturing inter-modal interactions. The coordinated feature representation space can be expressed as

$$f(x_i) \sim c(x_i) \sim v(x_i), \tag{6.1}$$

where x_i is the modalities; functions f, v, and c are the independent unimodal learning networks corresponding to FL, SER, and the baseline churn model, respectively, all within the coordinated representation space; and $\sim$ indicates similarity or coordination in this space's projection.

A considerable segment of multimodal machine learning focuses on translating or mapping between modalities. This involves producing a corresponding entity in a different modality from one that exists in another. For example, the aim could be to recognize and quantify various emotions from an audio signal, or to generate a corresponding score in a text based online survey. Multimodal translation issues have been extensively explored, with pioneering research encompassing speech synthesis, generating visual speech, video description, and cross-modal retrieval [164–167].

A translation dictionary was employed to map features from each modality, ensuring similarity and relevant patterns were preserved. The mapping matrix was articulated through various logical propositions, determined by domain experts, each corresponding to specific condition(s) on the unimodal features. Outputs from each unimodal data structure were transposed into two-level numeric nominal variables, based on a predefined prediction confidence $P(\text{pred})$ threshold, establishing a coordinated representation space via logical propositions. Constant weights were assigned to each matrix array C_i, F_i, and V_i, reflective of the unimodal feature's weight in multimodal learning, expressed as

$$\forall x_i, P1(x_i) : f(x_i) < P(\text{pred}) \implies F_i = 1; \neg P1(x_i) \implies F_i = 0$$

$$\forall x_i, P2(x_i) : c(x_i) \leq P(\text{pred}) \implies C_i = 0; \neg P2(x_i) \implies C_i = 2$$

$$\begin{aligned} &\forall x_i, P3(x_i) : v(x_i) \in \{\text{'Happiness','Neutral'}\} \implies V_i = 0; \\ &\neg P3(x_i) : v(x_i) \in \{\text{'Sadness','Anger'}\} \implies V_i = 1, \end{aligned}$$

where $P1(x_i)$, $P2(x_i)$, and $P3(x_i)$ correspond to the propositions associated with each unimodal model output.

Allocating appropriate weights for each feature indicator presents a complex endeavor. The proposed methodology assigned weights to the indicators by utilizing insights from experts with industrial domain knowledge, e.g. superannuation fund industry. This provided subtle alignment of feature translation, a crucial intermediary phase in data mapping and essential for any subsequent analytical procedure.

6.3.5 Hybrid fusion strategy

Based on the experimental study of different multimodal and fusion frameworks, a strategy was selected that employed a hybrid approach drawing on complementary insights from unimodal model outputs while preserving logical consistency. This incorporated late and decision level fusion, creating a multi-level fusion framework. Constant fusion weight $C_i + F_i + V_i \geq 0$ and $F_i = V_i$ was enforced and fusion weights for FL and SER models were assumed to have equal influential contributions to enhance baseline model performance. Therefore, a decision fusion D_i was defined to integrate complementary information and rank churn risk (low, mid, or high),

$$D_i = \begin{bmatrix} I_{\text{lowrisk}} \\ I_{\text{midrisk}} \\ I_{\text{highrisk}} \end{bmatrix} = I * \begin{bmatrix} C_1 + F_1 + V_1 \\ C_2 + F_2 + V_2 \\ \vdots \\ C_i + F_i + V_i \end{bmatrix} \tag{6.2}$$

where $D_i \in \{0, 4\}$ is the fusion output weight that represents the risk rank for each member i; C_i is the baseline customer churn model prediction outcome from historical data; F_i is the baseline customer churn model member's FL level; V_i is a qualitative indicator of the member's emotional disposition and satisfaction, detected from telephone interactions with the call center; $I(\cdot)$ is an indicator, where $I(\cdot) = 1$ if a specified condition holds true, and 0 otherwise. These relationships are subject to fusion conditions as follows.

- **Low risk churner.**

  $$\begin{aligned} I_{\text{lowrisk}} = & I(D_i = 0) \cdot I(C_i = 0) \cdot I(F_i = 0) \cdot I(V_i = 0) \\ & + I(D_i = 1) \cdot I(C_i = 0) \cdot I(F_i = 1) \cdot I(V_i = 0) \\ & + I(D_i = 1) \cdot I(C_i = 0) \cdot I(F_i = 0) \cdot I(V_i = 1) \end{aligned}$$

  If $I_{\text{lowrisk}} = 1$, customer is classified as low risk.

- **Mid risk churner.**

  $$\begin{aligned} I_{\text{midrisk}} = & I(D_i = 2) \cdot I(C_i = 2) \cdot I(F_i = 0) \cdot I(V_i = 0) \\ & + I(D_i = 2) \cdot I(C_i = 0) \cdot I(F_i = 1) \cdot I(V_i = 1) \end{aligned}$$

  If $I_{\text{midrisk}} = 1$, customer is classified as mid risk.

- **High risk churner.**

$$\begin{aligned}I_{\text{highrisk}} &= I(D_i = 4)\cdot I(C_i = 2)\cdot I(F_i = 1)\cdot I(V_i = 1)\\ &+I(D_i = 3)\cdot I(C_i = 2)\cdot I(F_i = 1)\cdot I(V_i = 0)\\ &+I(D_i = 3)\cdot I(C_i = 2)\cdot I(F_i = 0)\cdot I(V_i = 1)\end{aligned}$$

  If $I_{\text{highrisk}} = 1$, customer is classified as high risk.

Hence, the logical operators are structured to ensure that only one of I_{lowrisk}, I_{midrisk}, or I_{highrisk} is 1, i.e., true, for a particular member or customer risk level query, guaranteeing that each customer is exclusively assigned to one risk category. Combining these modalities generates a multimodal co-learning environment that presents a coordinated representation where the FL and SER modalities jointly enhance one another's training. Consequently, this co-learning approach overcomes drawbacks from depending solely on one data type and thus mitigates model bias.

Figure 6.3 shows that the proposed hybrid fusion framework includes three independent unimodal models, a coordinated feature representation space, feature mapping, and decision fusion mechanisms. The primary contribution is introducing a hybrid fusion approach to combine various data inputs into a unified feature representation space, effectively addressing heterogeneity across different modalities. This strategy is aimed at mitigating model bias. The critical challenge addressed here was formulating a hybrid fusion method that synergistically combined LF and decision fusion aspects, categorizing specific member churn into low, medium, or high risk.

The process to fuse knowledge from each modality using the proposed hybrid fusion learning approach comprises four steps as follows:

1. Map heterogeneous unimodal models $f(x)$, $c(x)$, and $v(x)$ data onto two-level numeric variables in a coordinated feature representation space (FRS).

2. Fused the unimodal feature values obtained from $f(x)$ and $v(x)$ and transfer them to $c(x)$ using LF.

3. Assign constant weights to each mapped modality feature value using logical propositions $P_n(x_i)$ to maintain pattern integrity.

4. Utilized the decision fusion matrix D_i to merge the weighted unimodal features, categorizing churn risk as low, mid, or high.

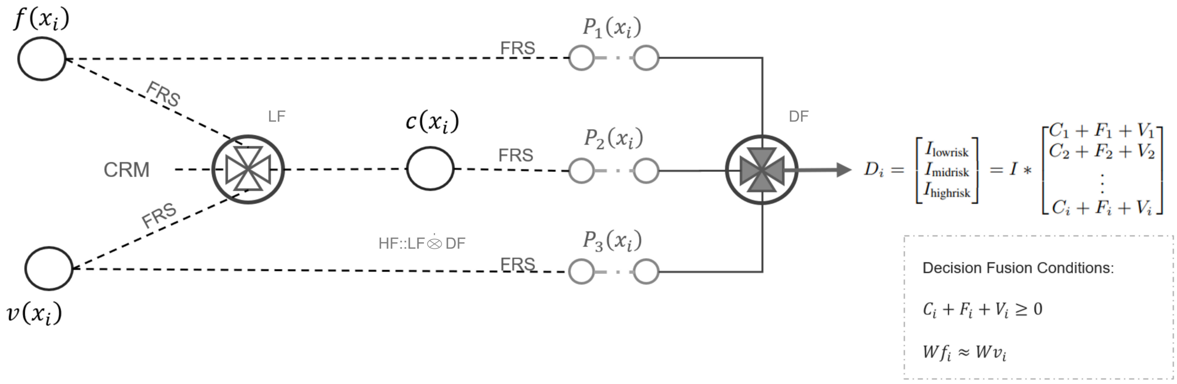

$f(x_i)$
FRS
FRS
$P_1(x_i)$
LF
$c(x_i)$
DF
$P_2(x_i)$
CRM
FRS
FRS
$D_i = \begin{bmatrix} I_{\text{lowrisk}} \\ I_{\text{midrisk}} \\ I_{\text{highrisk}} \end{bmatrix} = I * \begin{bmatrix} C_1 + F_1 + V_1 \\ C_2 + F_2 + V_2 \\ \vdots \\ C_i + F_i + V_i \end{bmatrix}$
HF::LF$\dot{\otimes}$DF
$P_3(x_i)$
FRS
$v(x_i)$
Decision Fusion Conditions:
$C_i + F_i + V_i \geq 0$
$Wf_i \approx Wv_i$

Figure 6.3: Proposed hybrid fusion process

The proposed hybrid fusion approach offers several advantages that are particularly beneficial for complex analytical tasks, such as customer or member behavior analysis in financial organizations.

1. Hybrid fusion provides a comprehensive analysis by capturing relationships at both low and high levels within the data, allowing for more in-depth understanding of underlying patterns. Using multiple data modalities also enhances prediction accuracy, since it leverages a broader range of information compared to single-modality models.

2. Hybrid fusion contributes to the model robustness by reducing overfitting risk through integrating EF, LF, and DF fusion techniques.

3. The hybrid fusion method provides a holistic view of customer and member behavior, offering 360-degree insights that enable more informed decision making based on a complete map of customer interactions and preferences.

These advantages make hybrid fusion an optimal choice for researchers and practitioners seeking to improve their predictive analytics efficacy.

6.4 Experiment Outcomes

6.4.1 Materials

The material for this study involved several real-world data sources, including financial transactions, CRM demographic details, and customer voice recordings. As mentioned in section 6.3.1, the privacy-preserving de-identification method was applied to the voice data, with emotional nuances retained by substituting original samples with analogous entries from established emotion databases. Moreover, datasets were carefully aligned based on customer behavior metrics from CRM and financial literacy scores, ensuring the model effectively matched customer records across different data modalities, crucial for reducing bias and enhancing multimodal learning outcomes.

6.4.2 Evaluation metrics for multimodal modeling performance

The mean average precision (MAP) metric was utilized for an objective assessment of the proposed method's effectiveness in ranking churn quality,

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{m_q}\sum_{k=1}^{n}P(k)\cdot rel(k), \tag{6.3}$$

where Q is the total number of queries; q is the specific query under consideration; m_q is the count of relevant churn risks for the q-th query; $P(k)$ is the precision at the k-th cutoff in the list; and $rel(k)$ an indicator function, where $rel(k) = 1$ signifies that churn risk at the k-th rank is relevant, otherwise $rel(k) = 0$.

Thus, MAP is computed by averaging precision for all instances across different risk levels (low, mid, high), with MAP = 1 indicates perfect system performance to identify churn risks. Average precision (AP) for individual risk rank queries is determined by the precision of each relevant risk rank retrieved and averaged across all queries to measure overall performance. The macro-averaged F1 score (MA-F1) was employed to address imbalances commonly found in the datasets. MA-F1 equally considers smaller and larger class performances, computed independently for each class as

$$\text{Macro-Averaged F1 Score} = \frac{1}{N}\sum_{i=1}^{N}F1\ \text{Score}_i, \tag{6.4}$$

where $F1$ Score$_i$ is the F1 score for the i-th class, and N is the total number of classes.

Table 6.2 shows an example case to illustrate the MAP evaluation metric, simulating churn risk prediction outcomes from a multimodal system.

Retrieved items are categorized as {low, mid, high}, and MAP is derived by computing the mean of the precision scores for each relevant item. In an ideal scenario, MAP = 1.0 indicates that the system can accurately identify all churn risk levels. The MAP calculation for the example data (six members) in Table6.2 is $(1+0.5+0.33+0.5+0.6+0.5)/6 = 0.57$, which suggests that, on average, system's predictions are 57% accurate at each cut-off level in the list. MAP is a crucial metric since it considers predictions

Table 6.2: Example evaluating churn risk prediction using MAP metrics

Member ID	**Risk (Actual)**	**Predicted Churn risk**	**Relevant (1, 0)** $rel(K)$	**Cumulative Relevant**	**Precision** $@K$**:** $P(K)$
1	Mid	Mid	1	1	1/1=1.0
2	High	Mid	0	1	1/2=0.50
3	Low	Mid	0	1	1/3=0.33
4	Low	Low	1	2	2/4=0.5
5	High	High	1	3	3/5=0.6
6	Mid	Low	0	3	3/6=0.5

precision across all churn risk levels, offering insights into the system's overall ability to member's rank churn risk accurately according to their likelihood of churning. The efficiency of each unimodal model within the framework was also assessed using other standard evaluation metrics, including the area under the curve (AUC), test accuracy, and recall. Combining these metrics provides a comprehensive performance evaluation for the effectiveness of each individual unimodal model in the overall multimodal system.

6.5 Results and Discussion

The empirical findings support the effectiveness of using multimodal data to enhance predictive accuracy, validating the thesis that a multidimensional approach to data analysis yields more reliable and actionable insights. By leveraging the strengths of each individual data modality, the hybrid fusion model not only improves the accuracy of churn predictions but also enriches the strategic decision-making process, providing a robust framework for enhancing customer retention and satisfaction.

6.5.1 Results

The proposed method was thoroughly evaluated regarding how integrating modality influenced overall performance, employing various evaluation metrics, including test accuracy, recall, F1 score, AUC, MAP, and MA-F1. Figure 6.4(a) shows the highest level of performance, achieving 91.2% test accuracy, was the hybrid fusion method utilizing the collective strengths of all modalities, namely FL, Churn, and SER. FL has the most pronounced influence on enhancing performance, surpassing that of SER. Figure 6.4(b) shows that implementing hybrid fusion optimized churn risk categorization, effectively reclassifies customers from low to mid and high risk categories. This confirms the hybrid fusion strategy robustness and capacity to provide a more nuanced analysis of churn risk, which is pivotal for strategic customer retention efforts.

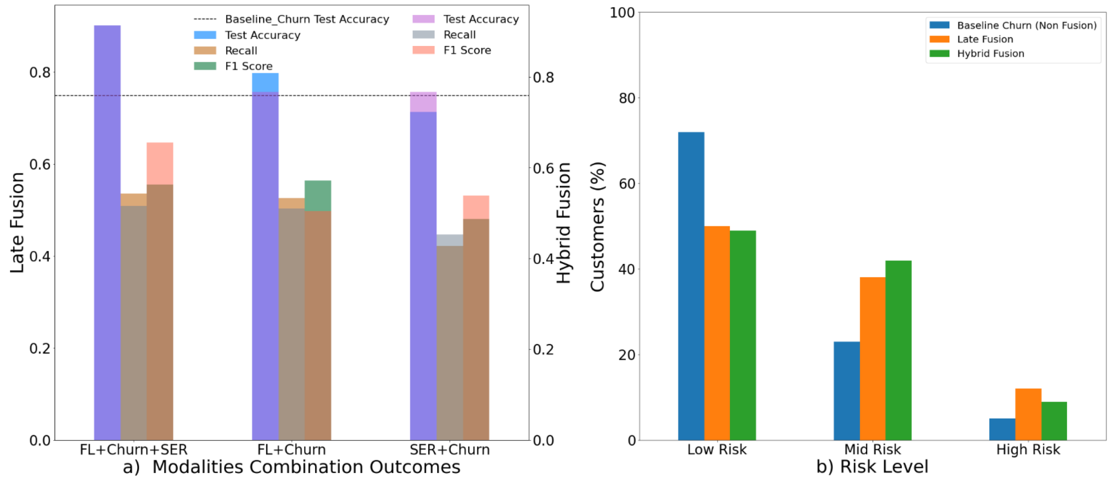


Figure 6.4: (a) Multimodal model performance significantly improved by combining LF and DF fusion (i.e., hybrid fusion); and (b) more members were identified as mid and high risk using the proposed hybrid fusion method.

Figure 6.5 compares ROC curves for late fusion and hybrid fusion. The higher AUC (Fig. 6.5(b)) confirms the advantages of combining multiple modalities and utilizing multi-level fusion benefits.

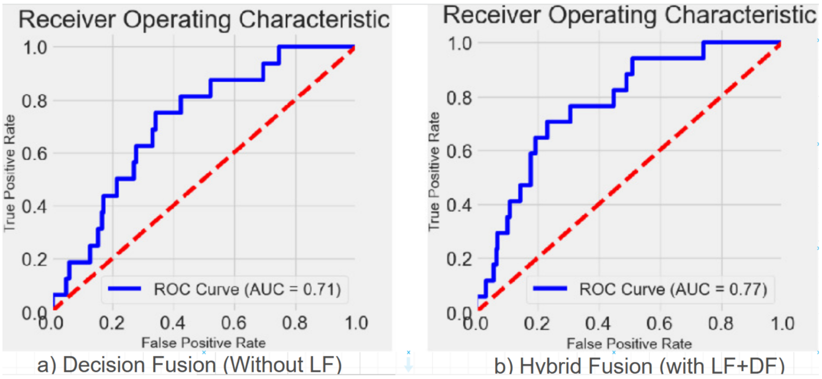


Figure 6.5: Hybrid fusion achieves higher AUC than the other methods considered

Table 6.3 compares risk identification accuracy for MAP and MA-F1 metrics. The multimodal learning approach, incorporating the hybrid fusion strategy, achieved significantly enhanced outcomes; MAP = 66 and MA-F1 = 54. This substantial improvement compared with the other strategies confirms hybrid fusion superior capability to accurately identify and categorize churn risk.

Table 6.3: Fusion method performance metrics

Metrics[a]	**Fusion method**		
	None	**DF (excludes LF)**	**hybrid fusion (LF+DF)**
MAP % ± STD	51 ± 0.8	65 ± 0.7	66 ± 0.1
MA F1 % ± STD	47 ± 0.1	47 ± 1.1	54 ± 0.6

[a] Higher value implies superior result

6.5.2 Discussion

This thesis presented a multimodal fusion learning framework that synergistically integrates customer's voice (CV), financial literacy (FL) survey data, and CRM records to predict churn risk across three categorizations: low, mid, and high. This multimodal approach marks the first attempt to include a multimodal hybrid fusion model to capture churn triggers better in a dynamic domain.

The SMOGN-COREG supervised model was used for the FL modality to extract FL levels from extensive unlabeled financial network data and questionnaire based survey data regarding financial behavior among members. The proposed multimodal model was critical to identify customers at higher churn risk due to inadequate FL. The baseline churn model leveraged a combination of SMOTE and ensemble ANN algorithms, achieving remarkable prediction accuracy for churn from large-scale, high-dimensional data. The SER model, which exploited the Mel Spectrogram components and a pre-trained CNN-VGG16, was instrumental in decoding emotional cues from member vocal interactions, adding a significant layer of behavioral insight.

A key empirical finding was the significant correlation between negative emotions and low FL with increased churn risk, identifying the psychological underpinnings for customer retention challenges. Comparing performance for the different modalities confirmed the distinct advantage from the hybrid fusion technique, achieving MAP = 66 and test accuracy = 91.2%, signaling its superiority to both non-fusion (single input model) and multimodal LF methodologies.

Despite these promising outcomes, some limitations remain for the proposed approach. The coordinated representation within multimodal fusion may fail to capture the rich intermodal information. Therefore, future study will explore joint representation strategies that can concatenate features from various modalities at the outset of the learning process.

Future study will develop the multimodal modal input spectrum by incorporating textual features as a fourth modality into the coordinated representation space, as shown in Figure 6.6. This addition will greatly enrich the analytical framework, allowing for more comprehensive analytic thinking about organizational member behavior. This approach excels in extracting semantic meaning and sentiment from text, which can provide additional layers of insight into customer satisfaction and intentions.

For example, in the context of the proposed framework, textual analysis could be aligned with data from customer voice (CV) and financial literacy (FL) assessments to provide a more nuanced view of customer emotions and potential churn triggers. Textual data could help clarify ambiguous vocal expressions or provide additional context to financial behaviors, enhancing the prediction accuracy of the churn model.

This thesis has contributed a novel framework for churn prediction and opened avenues for future innovations in multimodal learning. A step towards more empathetic, human-centric models that reflect customer's complex decision making processes of customers in the financial domain is to fuse diverse data types.

Python code for the proposed framework and further result visualization is available on the GitHub repository detailed in the footnote[1]. This resource simplifies reproduction and enhancement of the study's experimental results.

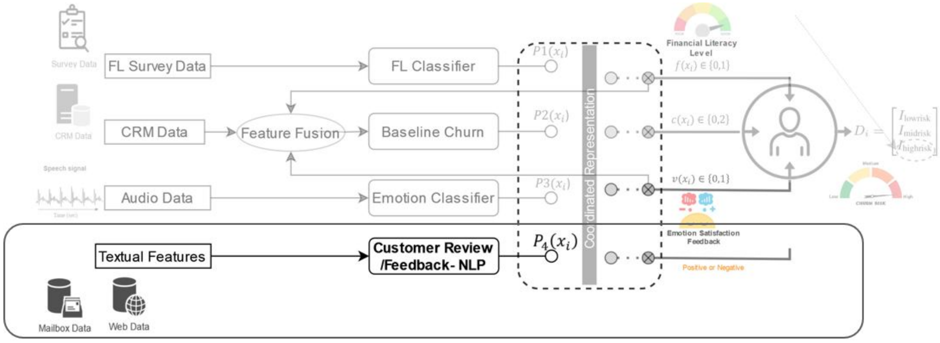


Figure 6.6: Incorporating textual features to represent member review and feedback from webpages and emails

[1] https://github.com/DavidHason/multimodal_churn_model

6.6 Summary

This chapter discussed is the concept for the thesis: employing distinct data sources to represent organizational member engagement for better churn prediction. The proposed hybrid fusion approach improved churn prediction in financial institutions by fusing emotional feedback from audio calls, historical CRM data, and FL levels. The main contributions from this study can be summarized as follows.

1. An innovative multimodal hybrid fusion framework that integrates CV, FL, and CRM data, offering an unprecedented approach to churn prediction.
2. Comprehensive analysis of customer behavior utilizing advanced machine learning techniques to extract insights from diverse data modalities, resulting in more accurate churn risk prediction.
3. Empirical validation of the correlation between emotional feedback from voice data, FL scores, and churn propensity, enhancing understanding of customer retention factors,
4. Confirmed superiority of the hybrid fusion model over traditional single-modality models using meaningful evaluation metrics, including mean average precision (MAP) and macro-averaged F1 Score (MA-F1), which confirm its effectiveness in predicting churn.

This study bridges the gap in churn prediction methods and sets the stage for a new era of customer retention strategies in financial organizations. This study contributes greatly to human-centric analytical model development and the innovative methodologies introduced will assist future studies to extend these models by incorporating additional modalities and exploring the potential for deep learning algorithms to further enhance prediction accuracy and understanding for member behavior.

Chapter 7

Chapter 7

7.1 Contributions

This thesis presents a comprehensive overview of quantitative research to analyze member engagement and churn in financial organizations and professional associations. Data mining offers various widely utilized methodologies to evaluate member engagement, which is crucial for maintaining customer relationships, effectively managing human resources, and making informed decisions.

Current methods cannot accommodate various situations due to the growing desire to explore more in-depth insights into member behavior.

Recent studies have explored useful information to analyze member behavior for churn. However, they often overlooked the importance of bridging the gap between member engagement, behavior, and churn through a holistic view of member's interactions, emotions, FL, and CRM. Strategies employing single data sources to address these issues have limitations, e.g. transactional, demographic, and textual data that do not provide a comprehensive image of member engagement and behavior. Therefore, this thesis investigated and proposed an innovative multidimensional data mining approach toward organizational member engagement, intending to capture the diverse insights required to bridge the current gap. Therefore, this thesis focused on the following aspects.

1. Chapter 1 described the overall thesis structure. Motivation and challenges for the previous studies were illustrated as essential issues for customer churn analysis

from a financial organization perspective. Several objectives were identified, and various churn propensity models to address the identified issues were proposed, evaluated, and discussed (including their limitations).

2. Chapter 2 examined member engagement and highlighted the pivotal role for FL to influencing member loyalty to an organization. Consequently, critical examination of member emotion recognition and its progressive methodologies emphasized the need to decode emotional influences on churn, emphasizing the essential impact from FL on informed decision-making, linking FL to member retention, and recommending a comprehensive churn prediction model that integrates an array of member engagement dimensions.

3. Chapter 3 explored speech emotion recognition (SER) approaches. SER significantly influences member engagement and churn within organizations. Although SER frameworks have evolved with integrating CNNs, the full potential for Mel spectrogram components as CNN inputs remains to be fully explored. Previous studies have yet to fully explore the application of variational mode decomposition in speech signal processing for emotion recognition, indicating a knowledge gap in the field. Chapter 3 presented the pioneering insight to use VMD for dynamic data augmentation in SER, introducing an innovative hybrid acoustic feature map technique that employs the CNN-VGG16 model for emotion extraction from speech signals. This marks the first approach of this type, and validates the model through empirical experiments. Combining prosodic and acoustic features enhanced SER model generalization, achieving state-of-the-art results, and setting new benchmarks for feature extraction and classification in emotion recognition from speech.

4. Chapter 4 investigated a significant step toward predicting financial literacy and has implications for member churn. Financial literacy is fundamental to enhancing member engagement within financial organizations, where informed decisions correlate with member satisfaction and churn. However, previously proposed (and implemented) methods, such as surveys and supervised learning, are limited by their reliance on labeled data and often overlook the multifaceted factors influencing FL, often hidden within unlabeled recorded data in financial network platforms and socio-economic status data. This oversight constrains the depth of FL analysis and its predictive accuracy for member churn. To address these limitations, this thesis proposed the SMOGN-COREG model, an innovative semi-supervised regression framework that exploits unbalanced and largely unlabeled financial

datasets. The proposed method significantly improved the model's predictive power. Pioneering the use of semi-supervised learning for FL prediction, the new model enhanced the FL performance accuracy, successfully labeling 64% of previously unlabeled data.

5. Chapter 5 considered how the significance of causal inference versus traditional churn prediction methods lies in not just predicting churn but understanding the reason behind it, enabling proactive member retention strategies. Several recent studies focused on leveraging causal Bayesian networks and counterfactual reasoning to explore deeper into churn triggers, but employing propensity score matching with DoWhy remains rarely considered for causal discovery, especially within financial organizations.

   Current studies often fail to link causality with churn predictions, a gap that motivated this thesis to develop an approach combining deep learning with PSM/-DoWhy for more robust analysis. This thesis also introduces a comprehensive churn propensity model incorporating SMOTE sampling, RFE, ensemble ANNs, and causal reasoning model, enhancing predictive accuracy and offering insights into churn root causes. Contributions include the first empirical investigation of causal Bayesian networks with PSM/DoWhy impacts on churn. This innovative methodology improves churn prediction and paves the way for applying counterfactual causal analysis, promising more profound insights into customer retention and behavior.

6. Chapter 6 considered how understanding member behavior, gauging engagement, and predicting churn could be achieved with multimodal modeling approaches. Previous studies almost exclusively utilized unimodal models with single data sources, e.g. textual features on social media, CRM data, transactional data, and demographic data separately for churn prediction. Traditional methods relying on singular data sources fall short of presenting an integrated view of member behavior, often failing to capture the dynamic nature of customer satisfaction and the detailed experiences behind churn.

   The proposed methods and models were motivated by the need to bridge the gap between member engagement and churn and address limitations for previous research that often overlooked the holistic nature of member interactions. This led to a multimodal hybrid fusion learning model combining various member engagement metrics, including CV, FL, and CRM data. The model explicitly considered their combined impact on member or client decisions and loyalty and offers an advanced churn prediction method.

Thus, this thesis aims to improve member retention strategies, helping to develop human-centric analytical models for the financial sector.

7.2 Future Work

Chapter 3 proposed a novel hybrid acoustic feature map technique that integrates harmonic and percussive components of Mel Spectrograms, utilizing the CNN-VGG16 model for advanced SER. The proposed second SER method is the first use of VMD for dynamic data augmentation in SER, significantly improving model generalization with notable test accuracy.

Future studies will develop this methodology by modifying the network architecture, with the intent to combine outputs from various neural networks, each trained on disparate acoustic features, creating an integrated model that captures a broader spectrum of emotional cues. Subsequently, including call transcripts, i.e., CV, as textual features will further reduce model bias in detecting emotion in different languages and improve model generalizability. Since the expressing emotions through vocal interaction varies across different cultures, the emotional acoustic features of speech signals do not directly match for different languages.

Model efficiency is also important, and future work will consider method(s) to optimize the VGG-optiVMD algorithm parameters, reducing computational demands while maintaining high accuracy, and exploring how best to include the most informative decomposed modes and their role(s) in acoustic feature extraction. Ultimately, this will identify a family of methods or frameworks to identify those decomposed modes that have informative emotional features, and creating an upper energy band filter will isolate those time frames of speech signals that carry significant emotional content. Initially, the speech signal will need to be separated into various related frames and the energy calculated for each frame to establish the median energy. This will then provide a suitable threshold, e.g. 50% of the median energy, to identify voiced frames. The identified voiced segments can then be arranged in a sequence to form a unified informative data frame. Finally, this voiced signal can be divided into overlapping frames and VMD applied to decompose only high-energy signals.

Chapter 4, introduced the SMOGN-COREG model, an innovative semi-supervised regression framework to analyze FL as a churn predictor. This model explicitly handled unbalanced and unlabeled datasets and integrated oversampling strategies with co-regression algorithms, significantly enhancing model predictive power by combining labeled and unlabeled data. The proposed SMOGN-COREG model significantly enhanced prediction accuracy, correctly labeling 64% of previously unlabeled data.

Future study will look to improve this model by incorporating additional data sources, such as behavioral and transactional records, to better understand member financial behaviors, which I have collectively called financial X-Ray. This will help develop more comprehensive framework(s) supporting organization member engagement strategies. There is also potential benefits from integrating SSL with other advanced data mining methods, to enhance financial behavior analysis, opening an exciting avenue for future work.

Chapter 5 proposed a new churn prediction technique by integrating causal analysis with machine learning techniques. Future study will expand this approach, first fine-tuning the proposed churn propensity model and employing smaller outcome windows to increase model sensitivity to ultimately detect instantaneous churn signals. Subsequent study will explore counterfactual causal analysis to deepen understanding of churn, offering a dynamic view of customer retention. The long-term goal is to introduce an advanced tool that can be embedded in existing CRM platforms to measure member engagement in real time and provide more insightful and actionable solutions for churn prevention.

Chapter 6 proposed an innovative multimodal hybrid fusion learning framework, a significant step forward in churn prediction methodologies. This framework combines CV, FL, and CRM databases to accurately determine churn risks. Future study will develop this framework further by addressing the now current limitations of current coordinated feature representation space methods. A suitable joint representation strategy could seamlessly concatenate features from various unimodal models before starting the learning process.

Incorporating textual features into the existing multimodal fusion framework can significantly enhance the model‚Äôs capacity to understand and predict customer behavior comprehensively. By utilizing large language models like GPT-4 or BERT, textual data from customer interactions such as emails, chat logs, or social media posts can be analyzed. These models are adept at extracting semantic meaning and sentiment, providing additional layers of insight into customer satisfaction and intentions. In the context of the proposed framework, textual analysis aligned with data from customer voice (CV) and financial literacy (FL) assessments can offer a nuanced view of customer emotions and potential churn triggers. Textual data can clarify ambiguous vocal expressions or provide additional context to financial behaviors, thereby enhancing the prediction accuracy of the churn model.

Further studies will explore incorporating textual features as a fourth unimodal model into the multimodal framework, a richer understanding of member behavior and engagement. This enhancement will create a more holistic picture of customer engagement, advancing a more human-centered approach in financial service organizations.

Extending the framework to other business sectors like education, where churn can manifest as student dropout or disengagement, could prove beneficial. This approach would integrate data from student interactions, academic performance, and textual feedback from evaluations. Future work will involve adapting the framework to include educational engagement metrics and communication logs, which could significantly improve retention and satisfaction in educational institutions. Such explorations validate the framework's versatility and enhance its robustness across various industry-specific challenges.

BIBLIOGRAPHY

[1] V. Kumar and W. Reinartz, *Customer relationship management*. Springer, 2018.

[2] Z. Chen and Z. Fan, “Building comprehensible customer churn prediction models: a multiple kernel support vector machines approach,” in *ICSSSM11*. IEEE, 2011, pp. 1–4.

[3] A. Lemmens and S. Gupta, “Managing churn to maximize profits,” *Marketing Science*, vol. 39, no. 5, pp. 956–973, 2020.

[4] J. Oechssler, A. Roider, and P. W. Schmitz, “Cognitive abilities and behavioral biases,” *Journal of Economic Behavior & Organization*, vol. 72, no. 1, pp. 147–152, 2009.

[5] L. F. Klapper, A. Lusardi, and G. A. Panos, “Financial literacy and the financial crisis,” National Bureau of Economic Research, Tech. Rep., 2012.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[7] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1d & 2d cnn lstm networks,” *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[8] D. H. Rudd, H. Huo, and G. Xu, “Predicting financial literacy via semi-supervised learning,” in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2022, pp. 304–319.

[9] H. Tran, N. Le, and V.-H. Nguyen, “Customer churn prediction in the banking sector using machine learning-based classification models.” *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 18, 2023.

[10] J. Spiess, Y. T'Joens, R. Dragnea, P. Spencer, and L. Philippart, "Using big data to improve customer experience and business performance," *Bell labs technical journal*, vol. 18, no. 4, pp. 3–17, 2014.

[11] European Parliament, "Artificial intelligence act: Meps adopt landmark law," https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law, 2024, accessed: 2024-08-07.

[12] A. Bharat, "Consumer engagement pattern analysis leading to improved churn analytics: an approach for telecom industry," in *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018, Volume 2*. Springer, 2019, pp. 203–211.

[13] H. O'ÄôBrien and P. Cairns, "Why engagement matters," *Cham: Springer International Publishing. doi*, vol. 10, pp. 978–3, 2016.

[14] J. S. Hastings, B. C. Madrian, and W. L. Skimmyhorn, "Financial literacy, financial education, and economic outcomes," *Annu. Rev. Econ.*, vol. 5, no. 1, pp. 347–373, 2013.

[15] J. Lamba and E. Jain, "Business revolution in post-covid era: an evolving economy outlook," in *Future of Work and Business in Covid-19 Era: Proceedings of IMC-2021*. Springer, 2022, pp. 19–30.

[16] S. H. Iranmanesh, M. Hamid, M. Bastan, G. Hamed Shakouri, and M. M. Nasiri, "Customer churn prediction using artificial neural network: An analytical crm application," in *Proceedings of the International Conference on Industrial Engineering and Operations Management, Pilsen, Czech Republic*, 2019, pp. 23–26.

[17] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[18] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*. IEEE, 2012, pp. 1–4.

[19] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, “Utterance-level multimodal sentiment analysis,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.

[20] Q. Jin, C. Li, S. Chen, and H. Wu, “Speech emotion recognition with acoustic and lexical features,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4749–4753.

[21] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[22] A. Milton, S. S. Roy, and S. T. Selvi, “Svm scheme for speech emotion recognition using mfcc feature,” *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[23] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using cnn,” in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[24] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[25] S. Demircan and H. Kahramanli, “Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech,” *Neural Computing and Applications*, vol. 29, pp. 59–66, 2018.

[26] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, “Speech emotion recognition using fourier parameters,” *IEEE Transactions on affective computing*, vol. 6, no. 1, pp. 69–75, 2015.

[27] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, “Emotion recognition in sound,” in *Advances in Neural Computation, Machine Learning, and Cognitive Research: Selected Papers from the XIX International Conference on Neuroinformatics, October 2-6, 2017, Moscow, Russia 19*. Springer, 2018, pp. 117–124.

[28] A. Satt, S. Rozenberg, R. Hoory *et al.*, “Efficient emotion recognition from speech using deep learning on spectrograms.” in *Interspeech*, 2017, pp. 1089–1093.

[29] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3d log-mel spectrograms with deep learning network,” *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[30] N. Hajarolasvadi and H. Demirel, “3d cnn-based speech emotion recognition using k-means clustering and spectrograms,” *Entropy*, vol. 21, no. 5, p. 479, 2019.

[31] L. S. Dendukuri and S. J. Hussain, “Emotional speech analysis and classification using variational mode decomposition,” *International Journal of Speech Technology*, vol. 25, no. 2, pp. 457–469, 2022.

[32] A. A. A. Zamil, S. Hasan, S. M. J. Baki, J. M. Adam, and I. Zaman, “Emotion detection from speech signals using voting mechanism on classified frames,” in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE, 2019, pp. 281–285.

[33] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, “Emotion recognition in sound,” in *International Conference on Neuroinformatics*. Springer, 2017, pp. 117–124.

[34] S. Kwon, “A cnn-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, vol. 20, no. 1, p. 183, 2019.

[35] A. M. Badshah, N. Rahim, and Ullah, “Deep features-based speech emotion recognition for smart affective services,” *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, 2019.

[36] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, pp. 101 894–101 904, 2020.

[37] D. H. Rudd, H. Huo, and G. Xu, “Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2022, pp. 392–404.

[38] G. J. Lal, E. Gopalakrishnan, and D. Govind, “Epoch estimation from emotional speech signals using variational mode decomposition,” *Circuits, Systems, and Signal Processing*, vol. 37, pp. 3245–3274, 2018.

[39] M. Zhang, B. Hu, X. Zheng, and T. Li, “A novel multidimensional feature extraction method based on vmd and wpd for emotion recognition,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 1216–1220.

[40] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis; using physiological signals,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[41] S. K. Khare and V. Bajaj, “An evolutionary optimized variational mode decomposition for emotion recognition,” *IEEE Sensors Journal*, vol. 21, no. 2, pp. 2035–2042, 2020.

[42] P. Pandey and K. Seeja, “Subject independent emotion recognition from eeg using vmd and deep learning,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 5, pp. 1730–1738, 2022.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[44] A. C. Worthington, “Predicting financial literacy in australia,” 2006.

[45] A. Lusardi, O. S. Mitchell, and V. Curto, “Financial literacy among the young,” *Journal of consumer affairs*, vol. 44, no. 2, pp. 358–380, 2010.

[46] R. Huang, H. Tawfik, M. Samy, and A. Nagar, “A financial literacy simulation model using neural networks: case study,” in *2007 Innovations in Information Technologies (IIT)*. IEEE, 2007, pp. 516–520.

[47] M. Ding, J. Tang, and J. Zhang, “Semi-supervised learning on graphs with generative adversarial nets,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 913–922.

[48] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” *Advances in neural information processing systems*, vol. 29, 2016.

[49] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[50] A. Tamaddoni, S. Stakhovych, and M. Ewing, “Comparing churn prediction techniques and assessing their performance: A contingent perspective,” *Journal of service research*, vol. 19, no. 2, pp. 123–141, 2016.

[51] A. Simion-Constantinescu, A. I. DAMIAN, N. ŢĂPUŞ, L.-G. PICIU, A. PURDILĂ, and B. Dumitrescu, “Deep neural pipeline for churn prediction,” in *2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet)*. IEEE, 2018, pp. 1–7.

[52] S. H. Dolatabadi and F. Keynia, “Designing of customer and employee churn prediction model based on data mining method and neural predictor,” in *2017 2nd international conference on computer and communication systems (ICCCS)*. IEEE, 2017, pp. 74–77.

[53] H. Lee, Y. Lee, H. Cho, K. Im, and Y. S. Kim, “Mining churning behaviors and developing retention strategies based on a partial least squares (pls) model,” *Decision Support Systems*, vol. 52, no. 1, pp. 207–216, 2011.

[54] V. Yeshwanth, V. V. Raj, and M. Saravanan, “Evolutionary churn prediction in mobile networks using hybrid learning,” in *Twenty-fourth international FLAIRS conference*, 2011.

[55] E. Domingos, B. Ojeme, and O. Daramola, “Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector,” *Computation*, vol. 9, no. 3, p. 34, 2021.

[56] A. Ghorbani, F. Taghiyareh, and C. Lucas, “The application of the locally linear model tree on customer churn prediction,” in *2009 International Conference of Soft Computing and Pattern Recognition*. IEEE, 2009, pp. 472–477.

[57] Y. Xie, X. Li, E. Ngai, and W. Ying, “Customer churn prediction using improved balanced random forests,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[58] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A churn prediction model using random forest: analysis of machine learning techniques

for churn prediction and factor identification in telecom sector," *IEEE access*, vol. 7, pp. 60 134–60 149, 2019.

[59] M. Shah, D. Adiga, S. Bhat, and V. Vyeth, "Prediction and causality analysis of churn using deep learning," *Comput. Sci. Inf. Technol*, vol. 9, no. 13, pp. 153–165, 2019.

[60] J. Pearl, "Causal inference in statistics: An overview," 2009.

[61] A. Bilal Zorić, "Predicting customer churn in banking industry using neural networks," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 14, no. 2, pp. 116–124, 2016.

[62] A. De Caigny, K. Coussement, K. W. De Bock, and S. Lessmann, "Incorporating textual information in customer churn prediction models based on a convolutional neural network," *International Journal of Forecasting*, vol. 36, no. 4, pp. 1563–1578, 2020.

[63] B. Culbert, B. Fu, J. Brownlow, C. Chu, Q. Meng, and G. Xu, "Customer churn prediction in superannuation: a sequential pattern mining approach," in *Databases Theory and Applications: 29th Australasian Database Conference, ADC 2018, Gold Coast, QLD, Australia, May 24-27, 2018, Proceedings 29*. Springer, 2018, pp. 123–134.

[64] A. Mishra and U. S. Reddy, "A novel approach for churn prediction using deep learning," in *2017 IEEE international conference on computational intelligence and computing research (ICCIC)*. IEEE, 2017, pp. 1–4.

[65] R. Mohan, S. Chaudhury, and B. Lall, "Temporal causal modelling on large volume enterprise data," *IEEE Transactions on Big Data*, vol. 8, no. 6, pp. 1678–1689, 2021.

[66] Y. Huang and M. Valtorta, "Identifiability in causal bayesian networks: A sound and complete algorithm," in *Proceedings of the national conference on artificial intelligence*, vol. 21, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 1149.

[67] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, "A survey on churn analysis in various business domains," *IEEE Access*, vol. 8, pp. 220 816–220 839, 2020.

[68] F. Lattimore and C. S. Ong, “A primer on causal analysis,” *arXiv preprint arXiv:1806.01488*, 2018.

[69] P. Gopal and N. B. MohdNawi, “A survey on customer churn prediction using machine learning and data mining techniques in e-commerce,” in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2021, pp. 1–8.

[70] N. N. Vo, S. Liu, J. Brownlow, C. Chu, B. Culbert, and G. Xu, “Client churn prediction with call log analysis,” in *Database Systems for Advanced Applications: 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part II 23*. Springer, 2018, pp. 752–763.

[71] T. Kimura, “Customer churn prediction with hybrid resampling and ensemble learning.” *Journal of Management Information & Decision Sciences*, vol. 25, no. 1, 2022.

[72] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, “Propension to customer churn in a financial institution: A machine learning approach,” *Neural Computing and Applications*, vol. 34, no. 14, pp. 11 751–11 768, 2022.

[73] R. Liu, S. Ali, S. F. Bilal, Z. Sakhawat, A. Imran, A. Almuhaimeed, A. Alzahrani, and G. Sun, “An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms,” *Applied Sciences*, vol. 12, no. 18, p. 9355, 2022.

[74] A. Guitart, P. P. Chen, and Á. Periáñez, “The winning solution to the ieee cig 2017 game data mining competition,” *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 252–264, 2018.

[75] J. T. Kristensen and P. Burelli, “Combining sequential and aggregated data for churn prediction in casual freemium games,” in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.

[76] O. Pierre-Yves, “The production and recognition of emotions in speech: features and algorithms,” *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[77] J. Q. Wang, T. Nicol, E. Skoe, M. Sams, and N. Kraus, "Emotion and the auditory brainstem response to speech," *Neuroscience letters*, vol. 469, no. 3, pp. 319–323, 2010.

[78] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[79] G. Aguilar, V. Rozgić, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," *arXiv preprint arXiv:1906.10198*, 2019.

[80] D. Alu, E. Zoltan, and I. C. Stoica, "Voice based emotion recognition with convolutional neural networks for companion robots," *Science and Technology*, vol. 20, no. 3, pp. 222–240, 2017.

[81] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language‚Äîa survey," *Emotion Recognition: A Pattern Analysis Approach*, pp. 237–267, 2015.

[82] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[83] J. Ahmad, M. Fiaz, S.-i. Kwon, M. Sodanil, B. Vo, and S. W. Baik, "Gender identification using mfcc for telephone applications-a comparative study," *arXiv preprint arXiv:1601.01577*, 2016.

[84] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[85] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Eleventh annual conference of the international speech communication association*, 2010.

[86] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[87] P. Motlıcek, “Feature extraction in speech coding and recognition,” Technical Report of PhD research internship in ASP Group, OGI-OHSU,< http ,Ä¶, Tech. Rep., 2002.

[88] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[89] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[90] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Interspeech 2014*, 2014.

[91] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends,” *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[92] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, “An image-based deep spectrum feature representation for the recognition of emotional speech,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484.

[93] D. Fitzgerald, “Harmonic/percussive separation using median filtering,” 2010.

[94] V. R. Carvalho, M. F. Moraes, A. P. Braga, and E. M. Mendes, “Evaluating five different adaptive decomposition methods for eeg signal seizure detection and classification,” *Biomedical Signal Processing and Control*, vol. 62, p. 102073, 2020.

[95] K. Dragomiretskiy and D. Zosso, “Variational mode decomposition,” *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2013.

[96] B. Basharirad and M. Moradhaseli, “Speech emotion recognition methods: A literature review,” in *AIP Conference Proceedings*, vol. 1891, no. 1. AIP Publishing LLC, 2017, p. 020105.

[97] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions,” *IEEE signal processing letters*, vol. 18, no. 2, pp. 130–133, 2010.

[98] U. Shrawankar and V. M. Thakare, “Techniques for feature extraction in speech recognition system: A comparative study,” *arXiv preprint arXiv:1305.1145*, 2013.

[99] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, “Feature learning in deep neural networks-studies on speech recognition tasks,” *arXiv preprint arXiv:1301.3605*, 2013.

[100] K. Aizawa, Y. Nakamura, and S. Satoh, *Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30-December 3, 2004, Proceedings, Part II.* Springer, 2004, vol. 3332.

[101] C. Harte, M. Sandler, and M. Gasser, “Detecting harmonic change in musical audio,” in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 21–26.

[102] J. L. Flanagan, *Speech analysis synthesis and perception.* Springer Science & Business Media, 2013, vol. 3.

[103] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[104] F. J. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[105] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[106] D. M. Schuller and B. W. Schuller, “A review on five recent and near-future developments in computational processing of emotion in the human voice,” *Emotion Review*, vol. 13, no. 1, pp. 44–50, 2021.

[107] S. Deb and S. Dandapat, “Fourier model based features for analysis and classification of out-of-breath speech,” *Speech Communication*, vol. 90, pp. 1–14, 2017.

[108] F. R. Kschischang, “The hilbert transform, the edward s,” *Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto*, vol. 12, 2006.

[109] M. R. Hestenes, “Multiplier and gradient methods,” *Journal of optimization theory and applications*, vol. 4, no. 5, pp. 303–320, 1969.

[110] R. T. Rockafellar, “A dual approach to solving nonlinear programming problems by unconstrained optimization,” *Mathematical programming*, vol. 5, no. 1, pp. 354–373, 1973.

[111] J. Lian, Z. Liu, H. Wang, and X. Dong, “Adaptive variational mode decomposition method for signal processing based on mode characteristic,” *Mechanical Systems and Signal Processing*, vol. 107, pp. 53–77, 2018.

[112] Z. Wang, G. He, W. Du, J. Zhou, X. Han, J. Wang, H. He, X. Guo, J. Wang, and Y. Kou, “Application of parameter optimized variational mode decomposition method in fault diagnosis of gearbox,” *Ieee Access*, vol. 7, pp. 44 871–44 882, 2019.

[113] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[114] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, “A database of german emotional speech.” in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[115] S. Schagen, A. Lines *et al.*, *Financial literacy in adult life: a report to the Natwest Group Charitable Trust*. NFER, 1996.

[116] C. Palm, “Measuring financial literacy of superannuation fund members: Preliminary results,” in *Proceedings of the 2015 Accounting and Finance Association of Australia and New Zealand (AFAANZ) Conference*. Accounting and Finance Association of Australia and New Zealand (AFAANZ), 2015, pp. 1–29.

[117] R. Behrman Jere, S. Mitchell Olivia, C. Soo, and D. Bravo, “Financial literacy, schooling, and wealth accumulation,” *American Economic Review*, vol. 102, pp. 300–304, 2012.

[118] P. Branco, L. Torgo, and R. P. Ribeiro, “Smogn: a pre-processing approach for imbalanced regression,” in *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR, 2017, pp. 36–50.

[119] G. Kostopoulos, S. Karlos, S. Kotsiantis, and O. Ragos, “Semi-supervised regression: A recent review,” *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 2, p. 1483,Äì1500. [Online]. Available: https://doi.org/10.3233/JIFS-169689

[120] Z.-H. Zhou and M. Li, “Semi-supervised regression with co-training,” in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, ser. IJCAI’05. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005, p. 908,Äì913.

[121] P. Branco, R. P. Ribeiro, and L. Torgo, “Ubl: an r package for utility-based learning,” *arXiv preprint arXiv:1604.08079*, 2016.

[122] P. Branco, L. Torgo, and R. P. Ribeiro, “SMOGN: a pre-processing approach for imbalanced regression,” in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, ser. Proceedings of Machine Learning Research, L. Torgo, B. Krawczyk, P. Branco, and N. Moniz, Eds., vol. 74. ECML-PKDD, Skopje, Macedonia: PMLR, 22 Sep 2017, pp. 36–50. [Online]. Available: http://proceedings.mlr.press/v74/branco17a.html

[123] S. Vluymans, *Learning from Imbalanced Data*. Cham: Springer International Publishing, 2019, pp. 81–110.

[124] L. Torgo and R. Ribeiro, “Utility-based regression,” in *Knowledge Discovery in Databases: PKDD 2007*, J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 597–604.

[125] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, “A multi-scheme semi-supervised regression approach,” *Pattern Recognition Letters*, vol. 125, pp. 758–765, 2019.

[126] R. H. Dehejia and S. Wahba, “Propensity score-matching methods for nonexperimental causal studies,” *Review of Economics and statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[127] A. Sharma and E. Kiciman, “Dowhy: An end-to-end library for causal inference,” *arXiv preprint arXiv:2011.04216*, 2020.

[128] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[129] M. A. Lejeune, “Measuring the impact of data mining on churn management,” *Internet Research*, vol. 11, no. 5, pp. 375–387, 2001.

[130] M. EconML, “Econml: A python package for ml-based heterogeneous treatment effects estimation,” 2019.

[131] D. Kalainathan, O. Goudet, and R. Dutta, “Causal discovery toolbox: Uncovering causal relationships in python,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 1406–1410, 2020.

[132] J. Runge, “Tigramite-causal discovery for time series datasets,” 2017.

[133] D. L. Donoho *et al.*, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS math challenges lecture*, vol. 1, no. 2000, p. 32, 2000.

[134] P. C. Kainen, “Utilizing geometric anomalies of high dimension: When complexity makes computation easier,” 1997.

[135] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[136] O. R. Zaïane, “Principles of knowledge discovery in databases,” *Department of Computing Science, University of Alberta*, vol. 20, 1999.

[137] A. H. Karp, “Using logistic regression to predict customer retention,” in *Proceedings of the Eleventh Northeast SAS Users Group Conference. http://www. lexjansen. com/nesug/nesug98/solu/p095. pdf*, vol. 15, 1998.

[138] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[139] O. Kramer, *Machine learning for evolution strategies*. Springer, 2016, vol. 20.

[140] R. K. Pedwell, J. A. Hardy, and S. L. Rowland, "Effective visual design and communication practices for research posters: Exemplars based on the theory and practice of multimedia learning and rhetoric," *Biochemistry and Molecular Biology Education*, vol. 45, no. 3, pp. 249–261, 2017.

[141] Y. C. Bilge, M. K. Yucel, R. G. Cinbis, N. Ikizler-Cinbis, and P. Duygulu, "Red carpet to fight club: Partially-supervised domain transfer for face recognition in violent videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3358–3369.

[142] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 800–10 809.

[143] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Clue: Cross-modal coherence modeling for caption generation," *arXiv preprint arXiv:2005.00908*, 2020.

[144] Y. Mao, Q. Sun, G. Liu, X. Wang, W. Gao, X. Li, and J. Shen, "Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation," *arXiv preprint arXiv:2010.07637*, 2020.

[145] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber, "Compositional learning of image-text query for image retrieval," in *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 2021, pp. 1140–1149.

[146] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[147] E. Petajan and H. P. Graf, "Automatic lipreading research: Historic overview and current work," in *Multimedia Communications and Video Coding*. Springer, 1996, pp. 265–275.

[148] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[149] D. Wang, P. Cui, M. Ou, and W. Zhu, “Deep multimodal hashing with orthogonal regularization,” in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[150] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, “Multi-view missing data completion,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1296–1309, 2018.

[151] L. Wang, W. Sun, Z. Zhao, and F. Su, “Modeling intra-and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval,” *Signal Processing*, vol. 131, pp. 249–260, 2017.

[152] H. Liu, F. Li, X. Xu, and F. Sun, “Multi-modal local receptive field extreme learning machine for object recognition,” *Neurocomputing*, vol. 277, pp. 4–11, 2018.

[153] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, “Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2321–2334, 2016.

[154] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[155] H. P. Martínez and G. N. Yannakakis, “Deep multimodal fusion: Combining discrete events and continuous signals,” in *Proceedings of the 16th International conference on multimodal interaction*, 2014, pp. 34–41.

[156] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, “Combining modality specific deep neural networks for emotion recognition in video,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 543–550.

[157] J. Ni, X. Ma, L. Xu, and J. Wang, “An image recognition method based on multiple bp neural networks fusion,” in *International Conference on Information Acquisition, 2004. Proceedings.* IEEE, 2004, pp. 323–326.

[158] M. Gönen and E. Alpaydın, “Multiple kernel learning algorithms,” *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[159] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[160] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.

[161] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[162] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” *Advances in neural information processing systems*, vol. 25, 2012.

[163] S. K. D’mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[164] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[165] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, “Text-to-visual speech synthesis based on parameter generation from hmm,” in *Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP’98 (cat. No. 98CH36181)*, vol. 6. IEEE, 1998, pp. 3745–3748.

[166] A. Kojima, T. Tamura, and K. Fukunaga, “Natural language description of human activities from video images based on concept hierarchy of actions,” *International Journal of Computer Vision*, vol. 50, pp. 171–184, 2002.

[167] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 251–260.