# ENHANCING EARLY DETECTION OF DEMENTIA USING INTER-RELATION-BASED FEATURES AND OVERSAMPLING TECHNIQUE

YANAWUT CHAIYO

DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2025

# ENHANCING EARLY DETECTION OF DEMENTIA USING INTER-RELATION-BASED FEATURES AND OVERSAMPLING TECHNIQUE

YANAWUT CHAIYO

THIS DISSERTATION IS A PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2025

# DISSERTATION APPROVAL
# MAE FAH LUANG UNIVERSITY
# FOR
# DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING

**Dissertation Title:** Enhancing Early Detection of Dementia Using Inter-Relation-Based Features and Oversampling Technique

**Author:** Yanawut Chaiyo

**Examination Committee:**

| | |
|---|---|
| Associate Professor Adisorn Leelasantitham, Ph. D. | Chairperson |
| Associate Professor Punnarumol Temdee, Ph. D. | Member |
| Associate Professor Roungsan Chaisricharoen, Ph. D. | Member |
| Associate Professor Nattapol Aunsri, Ph. D. | Member |
| Assistant Professor Chayapol Kamyod, Ph. D. | Member |

**Advisor:**

P. Temdee ..........................................................Advisor

(Associate Professor Punnarumol Temdee, Ph. D.)

**Dean:**

..........................................................................

(Assistant Professor Nacha Chondamrongkul, Ph. D.)

# ACKNOWLEDGEMENTS

The researcher would like to express the deepest gratitude to the Computer and Communication Engineering for Capacity Building Research Center, School of Applied Digital Technology, Mae Fah Luang University, Chiang Rai, Thailand, for their generous support and provision of research resources, which have been instrumental to the success of this study. I am also sincerely thankful to Chiang Rai Phachanukroh Hospital, Chiang Rai Province, Thailand, for their valued collaboration and the provision of critical data that served as the foundation of this study.

I am also grateful to Mae Fah Luang University for providing both academic and financial support during my studies, including a scholarship specifically granted for the completion of this doctoral dissertation.

The researcher is profoundly grateful to Associate Professor Dr. Punnarumol Temdee, my principal advisor, for her insightful guidance, academic rigor, and unwavering support throughout this work. In addition, I would like to express my heartfelt appreciation to the thesis examination committee members, Associate Professor Dr. Adisorn Leelasantitham, Associate Professor Dr. Roungsan Chaisricharoen, Associate Professor Dr. Nattapol Aunsri, and Assistant Professor Dr. Chayapol Kamyod for their invaluable insights, scholarly comments, and thoughtful suggestions, all of which significantly enriched the quality of this study.

Yanawut Chaiyo

**Dissertation Title** Enhancing Early Detection of Dementia Using Inter-Relation-Based Features and Oversampling Technique
**Author** Yanawut Chaiyo
**Degree** Doctor of Philosophy (Computer Engineering)
**Advisor** Associate Professor Punnarumol Temdee, Ph. D.

## ABSTRACT

Dementia affects both individuals and society, making early detection essential for effective management. However, reliance on advanced laboratory tests and specialized expertise limits accessibility, hindering timely diagnosis. To address this challenge, this study pioneers a novel approach by employing readily available biochemical and physiological features from electronic health records to develop a machine learning-based binary classification model, enhancing accessibility and early detection. This study utilizes a dataset from Phachanukroh Hospital in Chiang Rai, Thailand, for model construction. A hybrid data enrichment framework using feature augmentation and data balancing was proposed to increase data dimensionality. Inter-relation-based Features (IRFs) were suggested as a means to enhance data diversity and promote explainability by making features more informative through the application of medical domain knowledge. To balance the data, K-Means Synthetic Minority Oversampling Technique (K-Means SMOTE) was applied to generate synthetic samples in underrepresented regions of the feature space, improving class imbalance handling. Extra Trees (ET) was proposed for model construction because of its noise resilience and ability to manage multicollinearity. The performances were compared with Support Vector Machine (SVM), K-nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest (RF), and Gradient Boosting GB. Results revealed that the ET model significantly outperformed other models for the combined dataset with four Inter-Relation-Based Features (IRFs) and K-Means SMOTE across key metrics, including accuracy (96.47 %), precision (94.79 %), recall (97.86 %), F1-score (96.30%), and area under the curve of the Receiver Operating Characteristic (99.51 %).

**Keywords:** Dementia, Classification, Machine learning, Oversampling Technique

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Dementia has become a critical global issue, exacerbated by the aging population, leaving patients increasingly dependent and facing death (Castellazzi et al., 2020). Currently, 55 million people lived with dementia, a number projected to rise to 152.8 million by 2050 (Gustavsson et al., 2023; Nichols et al., 2022). The economic burden of dementia care exceeded $1 trillion in 2018 and was set to double by 2030, potentially surpassing $2 trillion as the global population ages and dementia cases rise (World Health Organization, 2019; Lastuka et al., 2024). In Thailand, the aging population is driving a steady increase in dementia patients, with numbers expected to grow by 10% annually (Muangpaisan, 2013; Thongwachira et al., 2019). Dementia significantly impairs daily activities, causing memory loss, confusion, and communication challenges, and often leads to complete dependency. Age is the primary risk factor, with most cases occurring in individuals over 65 years old, but lifestyle choices, cardiovascular health, and education also play a role. Beyond its impact on individuals, dementia has profound societal and economic consequences, with no cure and a slow progression. The increasing demand for care will strain healthcare systems, lead to higher costs, and prompt changes in societal structures, emphasizing the urgent need for policies addressing elder care and dementia management.

Dementia diagnosis involves a comprehensive assessment of symptoms and brain function. The detection of cognitive decline enables proactive preparation and the potential for behavior modification to mitigate the onset of dementia, as well as the identification of suitable treatment options. Typically, the diagnosis of dementia relies on a combination of physical examination, laboratory testing (such as brain function tests), and radiology, facilitated by advanced laboratory techniques (Gustavsson et al., 2023; Nichols et al., 2022). Traditionally, a detailed medical history, family history, and neurological examination are used to assess cognitive functions, including memory, thinking, and language. Standardized tests, such as the Mini-Mental State Examination (MMSE) or the Montreal Cognitive Assessment (MoCA), are generally used to measure cognitive abilities. Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI) are well-established methods for providing visual images of the brain to identify abnormalities, such as brain atrophy or tumors. Additionally, Positron Emission Tomography (PET) scans provide a more detailed view of brain function by tracking blood flow and metabolism. Furthermore, blood tests and biochemical analyses help rule out other potential causes of dementia symptoms, like hormonal imbalances or infections. Generally, dementia diagnosis is expensive, but it is necessary to have an advanced laboratory and a medical professional to ensure diagnosis accuracy, which is difficult to access for the public, especially in rural areas.

While traditional methods rely heavily on clinical evaluations and medical imaging, recent advancements in machine learning (ML) have introduced new possibilities for more accurate and efficient diagnosis of dementia. ML algorithms can analyze large datasets from medical images and psychological tests to aid in dementia diagnosis. They extract relevant features from brain images, like the degree of brain tissue loss or the enlargement of brain ventricles. They can analyze results from psychological tests to identify patterns associated with dementia (Gómez et al., 2017). By tracking health data and cognitive function over time, ML can predict the onset of dementia. Recently, deep learning has been trained to analyze MRI or PET images and identify subtle changes associated with dementia, such as brain atrophy or abnormal protein deposits. As mentioned earlier, the widespread adoption of ML has shown promising results in utilizing various types of data for dementia prediction. It has also been demonstrated that one of the challenges of using an ML-based model is the availability of high-quality and diverse datasets for training. Insufficient high-quality data from data collection could hinder model performance. Feature augmentation was one of the popular approaches, widely used for addressing the low dimension of the dataset (Gómez et al., 2017; Jeong et al., 2001; Nancy et al., 2017; Trambaiolli et al., 2017; Rodrigues et al., 2013; Pritchard et al., 1994). By capturing complex relationships that linear models might miss, feature augmentation can typically improve model performance. This method has been applied in several previous works to modify existing features within a dataset to improve ML models (Pritchard et al., 1994; Shorten et al., 2019).

While traditional approaches rely heavily on clinical evaluations and medical imaging, this study emphasizes the use of clinical data derived from electronic health records (EHRs) to improve the early detection of dementia, aiming to promote broader accessibility for the general population. More specifically, this study aims to propose a classification method that effectively classifies patients into those with and without dementia. Since the raw data from hospital was of low quality and low dimensionality, necessitating effective hybrid data enrichment method and making it suitable to improve model performance. In this study, new features, named Inter-Relation-Based Features (IRF), were constructed through a feature augmented process and integrated into the original dataset. Applying medical domain knowledge, IRF represents relationships between features within the same groups to provide more informative features and promote explainability. To tackle the data imbalance, K-means SMOTE was used to synthetically increase the number of minority class instances, which are patients with dementia. K-means SMOTE preserves data distribution, reduces bias from oversampling in noisy regions, and enhances model performance on imbalanced datasets. Extra Trees (ET) was proposed as an effective classifier due to its ability to handle multicollinearity and its potential to work with complex, noisy, or high-dimensional data. The proposed model was compared with other existing ML methods including Support Vector Machine (SVM), K-nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest (RF), and Gradient Boosting (GB). The models were evaluated using the confusion matrix, accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Receiver Operating Characteristic Curve (ROC).

## 1.2 Objectives

This study aims to address the limitations of traditional dementia diagnostic methods by applying machine learning to clinical data obtained from electronic health records (EHRs). The specific objectives are as follows:

1.2.1 Enhance dementia classification by constructing and integrating Inter-Relation- Based Features (IRFs) derived from medical domain knowledge, while addressing class imbalance using K-Means SMOTE to improve data distribution and overall model performance.

1.2.2 The proposed Extra Trees-based model is developed and evaluated in comparison with other machine learning algorithms including Support Vector Machine (SVM), K-nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest (RF), and Gradient Boosting (GB). based on diagnostic metrics such as confusion matrix, accuracy, precision, recall, F1-score, Area under the curve (AUC), and Receiver operating characteristic curve (ROC).

## 1.3 Scope of Research

This study focuses on the early screening of dementia using clinical data derived from electronic health records (EHRs), rather than relying on costly diagnostic methods such as neuroimaging or advanced laboratory testing. The scope of the data is limited to structured data that can be routinely collected in general healthcare settings, including blood test results, vital signs, and basic demographic information.

The study emphasizes the construction of new features, referred to as Inter-Relation-Based Features (IRFs), which are derived from externally validated theories and calculation formulas that align with the characteristics of the dataset used in this research. These features are designed to enhance classification performance and support clinically meaningful interpretation. In addition, the K-Means SMOTE technique was employed to address the issue of class imbalance, and the Extra Trees (ET) algorithm was selected as the primary classifier. The model's performance was compared with other commonly used machine learning algorithms, including SVM, KNN, ANN, RF, and GB.

This study does not include medical imaging data (MRI, CT, or PET scans) or unstructured data such as clinical notes. It is confined to tabular data formats suitable for practical application in healthcare systems, particularly in resource-limited settings.

# CHAPTER 2

# LITERATURE REVIEWS

## 2.1 Features for Dementia Classification Model

ML-based classification or prediction models employ different types of patient data for constructing the model, such as medical records, health history, behavioral patterns, and biological data. Studies have shown that datasets commonly used in ML models can be divided into three categories: unstructured data such as images, structured data such as records from a database, and hybrid structured data, which is the combination of unstructured and structured data. For unstructured data, several complex models have been implemented to process and analyze visual data (Ullah et al., 2018; Castellazzi et al., 2020). For examples, recent advancements in deep learning have shown significant promise in neuroimaging-based dementia detection, particularly for Alzheimer's disease (AD). Deep learning models have consistently outperformed traditional machine learning approaches in analyzing MRI scans for early diagnosis (Bansal et al., 2022). Studies utilizing the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset have demonstrated that Convolutional Neural Networks (CNNs) and transfer learning architectures such as InceptionV3 and ResNet can achieve high classification accuracy ranging from 91.8% to 95.2% (Bansal et al., 2022; Narmatha et al., 2021). Notably, Dense Convolutional Networks (DenseNet) have recently outperformed other CNN variants, reaching an accuracy of 96.1% in early AD classification tasks (Vardhini et al., 2024).

In terms of structured data groups, ML models operate on data that is organized in a tabular format, such as databases or spreadsheets. The standard dataset is EHR, containing different types of medical data, personal data, and behavioral data. For example, there was a study proposed on a classification model to classify mild cognitive impairment and cognitively normal subjects using itemized scores of three widely used standard neuropsychological tests, including Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) and MMSE (Almubark et al., 2020). For this study, four different ML models were studied: SVM, RF, GB, and AdaBoost. Furthermore, the adaptive synthetic sampling technique was employed to enhance the performance of the SVM-based model. It proposed novel feature extraction techniques, namely the Feature Extraction Battery, for classifying dementia (Javeed et al., 2023). A study employed a GB-based model for multiclass classification of heart failure, aortic stenosis, and dementia using the EHR data from the hospital. (Yongcharoenchaiyasit et al., 2023)

In terms of the hybrid data group, a study of a classification model for dementia based on MRI and clinical data was proposed. They incorporated univariate feature selection as a preprocessing step to filter features from MRI data (Lastuka et al., 2024). The Latest advancements in ML for AD detection and classification, concentrating on neuroimaging studies and some related clinical data, were examined. The techniques explored include SVM, RF, CNNs, and K-means (Mirzaei & Adeli, 2022; Mohammed et al., 2021). It can be seen that research studies in medical applications often involve the use of various types of data, including neuroimaging, protein sequences, speech data, electroencephalogram (EEG) signals, and magnetoencephalography (MEG) signals, as well as additional data like medical history and genetic information. However, working with such large datasets is not always practical in medical settings.

In this study, structured data, specifically biochemical and physiological features derived from EHRs, are primarily utilized to develop ML-based classification models to promote accessible and convenient early detection of dementia for the general public. This proposed work is motivated by the limitations of neuroimaging and brain function assessments, which typically require advanced medical facilities and specialized expertise that may not be readily available in rural or resource-limited settings.

## 2.2 Feature Engineering for Disease Prediction

Feature engineering is a crucial step in the ML pipeline, especially for disease prediction or classification tasks. Effective feature engineering can significantly enhance the predictive power of models in disease prediction. Feature extraction is a crucial part of feature engineering. It involves transforming raw data into a format that is more suitable for modeling and analysis. Several feature extraction methods have been proposed to enhance the diagnostic accuracy of disease prediction or classification models (Mohammed et al., 2021). The examination of the time-frequency representation and feature extraction-based model to distinguish EEG segments of control subjects from those of AD patients was conducted (Cura et al., 2022). A dementia classification based on speech analysis of casual talk during a clinical interview could utilize speech feature extraction to reduce the dimensionality of the speech dataset for the SVM model (Hanai et al., 2022). While feature extraction can significantly improve ML performance, it also presents challenges that can negatively impact results. Reducing data dimensionality can lead to information loss, thereby degrading model performance. Some feature extraction methods are computationally expensive, especially when working with large datasets, and may capture noise, which can lead to overfitting. To ensure reliable performance, it is essential to select extraction methods and conduct thorough validation and testing carefully.

Feature augmentation is also a method within feature engineering, typically used to enhance the generalization or performance of the model. It involves enhancing a dataset by adding new features derived from the existing ones. Augmented features can be used to capture complex relationships and patterns that may not be explicit in the original features. Moreover, it can enable models to become more robust against data unpredictability and noise, hence producing more accurate predictions. Recently, the health dataset features were found to facilitate the generation of new data through mathematical combinations. By providing more informative features, augmented datasets can lead to better model accuracy and precision. Many studies employed feature augmentation techniques to improve the performance of disease prediction and classification models. For example, an extensive review of data augmentation methods applicable in computer vision domains has been conducted. The study results found that data augmentation methods based on explicit transformation operations provide accurate and reliable performance improvements (Mumuni & Mumuni, 2022). In addition, a study on a transfer learning technique for detecting and classifying the severity of dementia using MRI scan images was proposed (Jha et al., 2022).

While extracted features can highlight relevant patterns, augmentation can introduce variability, helping models learn from a broader spectrum of data. Many models were using both techniques for disease prediction or classification. For example, a structured approach was examined, including preprocessing, dimensionality reduction using principal component analysis (PCA), dataset augmentation with a neural network, training, and evaluation of the dementia dataset using CNN-based classification (Jha et al., 2022). Moreover, the combination of PCA and CNNs for AD detection was a powerful approach. PCA reduced dimensionality and improved efficiency, while CNNs excel at extracting discriminative features from images (Reddy et al., 2023).

In this study, feature augmentation is selected to play a central role in enhancing the informativeness and interpretability of the input space. New inter-feature constructs are generated by leveraging established medical knowledge and clinically validated relational equations to produce more meaningful and explainable features. This domain-driven augmentation enhances model performance and fosters transparency and trustworthiness in clinical decision support systems.

## 2.3 Data Balancing for ML Model Construction

Data balancing methods are employed to address class imbalance in datasets, particularly in classification problems where one class is significantly underrepresented compared to others. The dominant class may influence the models, resulting in low performance for the minority class, despite achieving great accuracy. Three main categories define data balancing techniques:  oversampling, undersampling, and hybrid techniques. Synthetic Minority Oversampling Technique (SMOTE) has become well-known among the methods of data balancing in the domain of ML. It essentially balances the dataset and prevents the model from overfitting by generating a synthetic dataset of minority class instances from the existing data.  Many studies on the diagnosis and classification of dementia and AD have highlighted the general importance of the SMOTE method.  For example, a study using image-enhancing techniques in conjunction with SMOTE and deep learning to improve the accuracy of early AD detection (Samanta et al., 2023).

K-Means SMOTE is an enhanced version of SMOTE that utilizes clustering to generate more meaningful synthetic samples, thereby increasing its effectiveness in specific scenarios. It combines K-means clustering with SMOTE to improve classification performance. In addition, it differs from original SMOTE by considering the overall data distribution by using clustering, whereas SMOTE operates locally on minority class instances without considering the global structure. Focuses on areas of the feature space where the minority class is sparse, K-Means SMOTE tends to reduce the risk of generating noisy samples.  It is an effective technique for addressing class imbalance in various prediction tasks, highlighting the versatility and effectiveness of K-Means SMOTE in enhancing prediction performance across diverse domains, especially for medical datasets (Liu et al., 2023; Hairani et al., 2020). Additionally, Adaptive Synthetic Sampling (ADASYN) has also been widely used for data synthesis. It adaptively generates more synthetic data for minority class samples that are harder to learn, i.e., those surrounded by many majority class samples. While adaptive to sample difficulty by focusing on harder-to-learn instances, ADASYN may overemphasize borderline or noisy samples, leading to overfitting and reduced generalizability in real-world clinical applications, especially when class boundaries are fuzzy or error-prone (He et al., 2008).

However, K-Means SMOTE stands out among oversampling techniques for disease prediction by combining clustering with synthetic sampling, enabling more informed data generation. Unlike SMOTE and ADASYN, it avoids noisy regions and reinforces representative areas, enhancing class separability and reducing overlap with the majority class. This makes it particularly well-suited for imbalanced clinical datasets with subtle decision boundaries, which justifies its selection for this study.

## 2.4 Machine Learning Classification for Dementia Prediction

Research studies on ML methods for dementia prediction and diagnosis highlight three main groups: traditional classifiers, ensemble learning, and deep learning. Traditional models, such as Naive Bayes, Decision Trees, and SVM, are effective for smaller datasets or less complex features. Ensemble methods, such as RF, ET, and GB, combine multiple models to improve performance, robustness, and generalization. Deep learning models excel at analyzing high-dimensional data, such as medical images and genomic sequences, automatically learning relevant features without the need for extensive manual engineering. However, ensemble-based methods are often preferred over deep learning in medical applications due to the challenges of obtaining sufficient high-quality data, which can be both difficult and time-consuming.

Ensemble learning, which combines multiple models to improve prediction accuracy, has become a powerful tool in the diagnosis and prognosis of various diseases, including dementia. For example, a study utilized handwriting analysis for diagnosing neurodegenerative disorders like AD and Parkinson's disease (Ranjan et al.2022). Another study developed an ensemble model using Light Gradient Boosting, Categorical Boosting, and Adaptive Boosting for AD detection (Öcal, 2024). Similarly, an ensemble deep learning approach for disease prediction through metagenomics was proposed (Shen et al., 2023). Furthermore, a comparison between traditional and ensemble classifiers for the multiclass classification of heart failure, aortic stenosis, and dementia was investigated (Vardhini et al., 2024). Meanwhile, ensemble methods were applied to the Open Access Series of Imaging Studies dataset for dementia prediction (Goel et al., 2023). These studies consistently demonstrate that ensemble methods outperform traditional models.

Recent studies have explored the use of ET and decision tree ensembles for improving predictions in various domains, especially disease prediction (Shafique et al., 2019; Aashima et al., 2021). ET enhances randomness by also randomly selecting the split threshold, rather than calculating the optimal split point. This dual randomization, which incorporates both feature and threshold selection, introduces greater diversity among the trees, resulting in reduced variance and improved generalization, particularly in high-dimensional datasets (Hanczár et al., 2023). This is particularly beneficial when dealing with noisy or complex data. Compared to other ensemble methods, such as RF, ET further reduces variance by utilizing random thresholds, making it more effective for small datasets where feature engineering is crucial (Zhou & Feng, 2017). While GB sequential error correction risks overfitting on small datasets and requires extensive hyperparameter tuning, ET’s randomized splits and ensemble averaging provide more stable performance with minimal tuning (Hanczár et al., 2023; Zhou & Feng, 2017). Additionally, ET's bagging approach naturally mitigates class imbalance, whereas GB may amplify bias when synthetic oversampling introduces noise (Fernández et al., 2018; Narasimhan et al., 2021). Deep learning models, while powerful, often require large-scale data and extensive hyperparameter tuning, making them less practical for low-dimensional clinical datasets (Wen et al., 2020).

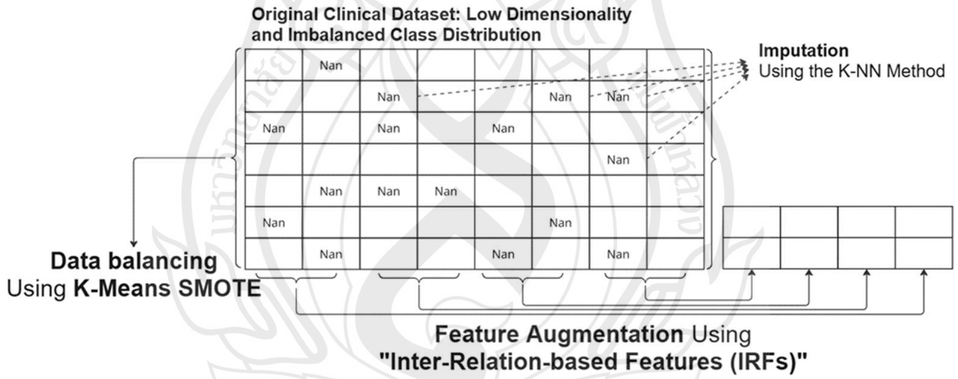
Within the realm of healthcare data, which often presents heterogeneous patterns, missing values, and irrelevant features, the added randomness of ET helps the model avoid overfitting to misleading correlations (Fernández-Delgado et al., 2014). Therefore, ET often outperforms traditional ensemble learning methods in biomedical applications, including disease classification, phenotyping, and risk prediction (Kourou et al., 2015). Its capacity to model complex, nonlinear interactions among features makes it particularly suitable for tasks like dementia classification in this study.

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Conceptual Framework

This study proposes a novel ML-based binary classification model using readily available biochemical and physiological features from electronic health records to improve accessibility and early detection. Since sufficiently high-quality data is a key to the success of ML-based models, this study proposes a hybrid data enrichment framework, feature augmentation, and data balancing to improve the overall classification performance of ML-based models. The conceptual diagram of the proposed hybrid enrichment framework is illustrated as follows.


Original Clinical Dataset: Low Dimensionality and Imbalanced Class Distribution
Imputation
Using the K-NN Method
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Nan
Data balancing
Using K-Means SMOTE
Feature Augmentation Using
"Inter-Relation-based Features (IRFs)"

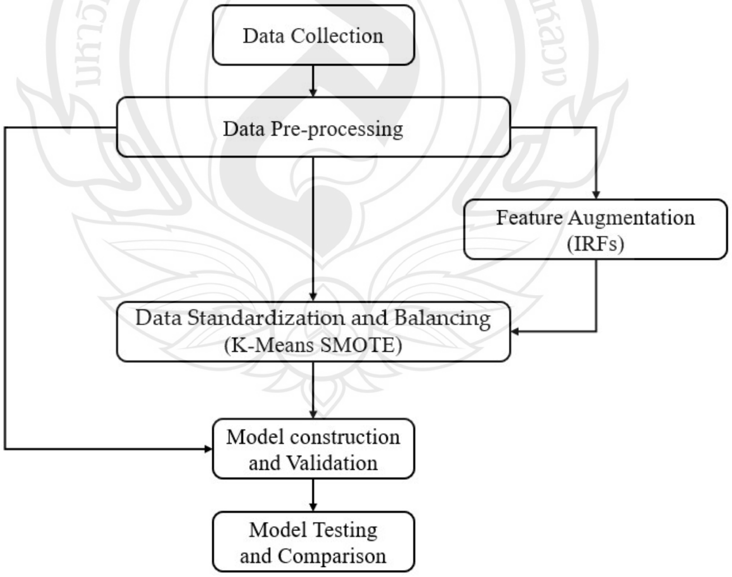
**Source** Chaiyo et al. (2025)

**Figure 3.1** Conceptual diagram of data enrichment framework, using Inter-Relation-Based Features for feature augmentation and K-Means SMOTE for data balancing

Figure 3.1 illustrates that the dimension of the original dataset increased for both the number of features and the number of examples using a hybrid data enrichment framework. For feature augmentation, the IRF was constructed from the features within the same group. In this study, there were five groups of features used for relationship presentation, including (1) blood pressure, (2) lipid levels, (3) blood sugar levels, (4) renal and chemical substances, and (5) blood cell count. The creation of new features based on the existing mathematical relationships, derived from medical domain knowledge, is expected to increase the diversity of the data without introducing new biases into the newly generated, more informative features and promote explainability. By creating interaction terms, relationships that might not be evident in the original features are expected to be captured. To balance the data, K-Means SMOTE was selected to increase the minority class, which was the patients with dementia class. Incorporating these augmented and synthesized new data alongside the original dataset was expected to improve the overall performance of the ML models. Furthermore, this study aims to determine which combination of datasets can enhance the ML-based classification models.

## 3.2 Methodology

The research methodology of this study is shown in Figure 3.2. Details of each process are discussed in this section.

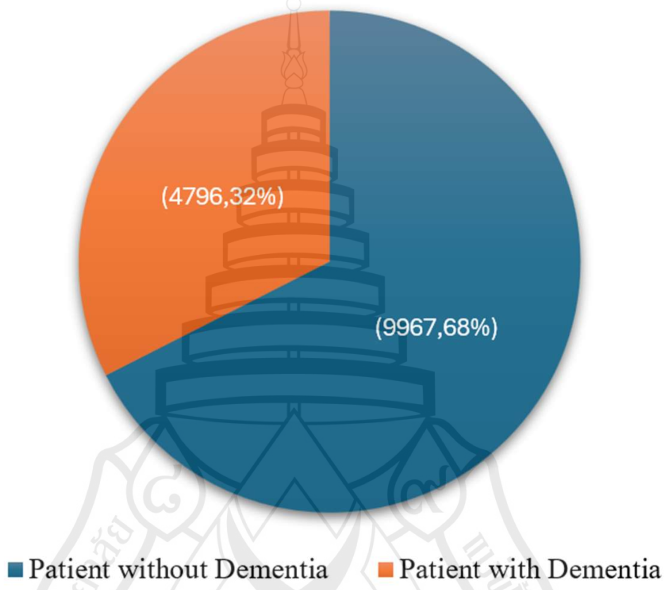


**Source** Chaiyo et al. (2025)

**Figure 3.2** Overview of the research methodology, encompassing data preparation, model development, and performance evaluation

### 3.2.1 Data Collection

This study used the EHR from Chiang Rai Phachanukroh Hospital, Chiang Rai province, Thailand, which comprises 14,763 records and 22 original features. There are 4,796 records of patients with dementia and 9,967 records of patients without dementia (heart failure and heart valve disorders). The data portion is shown in Figure 3.3.



**Source** Chaiyo et al. (2025)

**Figure 3.3** Class portion of the original dataset, having patients without dementia as the majority class and patients with dementia as the minority class

The feature datasets are divided into two categories: personal (e.g., Age, Height, Weight) and five other groups of clinical features, as shown in Table 3.1.

**Table 3.1** Original dataset and feature group

| No. | Group | Feature | Data Range |
|---|---|---|---|
| 1. | Personal Data | Age | 73.17±8.54 |
| | | Weight (W) | 56.37±9.70 |
| | | Height (H) | 155.03±7.62 |
| | | Gender (S) | 0-1 |
| 2. | Blood Pressure | Systolic Blood Pressure (SBP) | 130.00±16.34 |
| | | Diastolic Blood Pressure (DBP) | 68.93±9.9. |
| 3. | Lipid Levels | Cholesterol (Chol) | 171.02±28.77 |
| | | Triglyceride (TG) | 115.81±33.59 |
| | | Low-Density Lipoprotein (LDL) | 112.19±21.17 |
| | | High-Density Lipoprotein (HDL) | 44.95±8.67 |
| 4. | Blood Sugar Level | Fasting Blood Sugar (FBS) | 123.47±62.52 |
| 5. | Minerals and Chemical Substances | Creatinine (Cr) | 1.59±1.45 |
| | | Blood Urea Nitrogen (BUN) | 25.96±16.44 |
| | | Hemoglobin (Hb) | 11.34±1.72 |
| | | Potassium (K) | 3.96±0.49 |
| | | Sodium (Na) | 137.67±3.02 |
| 6. | Blood Cells | White Blood Cell (WBC) | 9006.04±2639.41 |
| | | Neutrophil (Neut) | 74.65±12.80 |
| | | Platelet (Plt) | 238667.17±60548.65 |
| | | Lymphocyte (Lymph) | 18.12±7.45 |

### 3.2.2 Data Preprocessing

Data preprocessing is a crucial step used in preparing data ready for ML. In this study, the data preprocessing process consists of two main steps, including data cleaning and imputation. Due to human error in this dataset, some features were left unnamed and could not be used. More specifically, features like education level and service date were considered irrelevant to the models' predictions. For data cleaning, any feature with over 90% missing data was excluded as it lacked sufficient information to be useful. For categorical features, one-hot encoding was employed. It involves transforming each category into a separate binary feature. In this study, the “gender” column was encoded as two binary features: one for “male” (represented as 1) and another for “female” (represented as 0).

To address the missing values, the K-NN imputation method was employed to maintain the relationships within the data during data imputation (Beebe-Wang et al., 2021). The K-NN imputation method was a technique used to handle missing data in a dataset by estimating the missing values based on the values of similar (neighboring) data points (Pujianto et al., 2019). The core idea is to find the most similar data points (nearest neighbors) to the data point with missing values and use them to predict the missing values. The K-NN imputation method specifies the parameter k, which represents the number of nearest neighbors used as references for imputing missing values. In this study, the value of k was set to 2 to emphasize imputation based on actual surrounding data points while minimizing reliance on averaged values that may obscure individual variability. When a missing value is encountered, the algorithm calculates the distance between the incomplete row and all other rows with complete data using Euclidean distance metrics. The two nearest neighbors are then selected as the basis for imputation. Additionally, the missing attribute is numerical, and the imputed value is calculated as the average of the corresponding values from the two selected neighbors. This imputed value is then substituted into the original row, replacing the missing entry, to ensure consistency with the most similar existing data points.

### 3.2.3 Feature Augmentation

The number of Inter-Relation-Based Features (IRFs) was designed based on the structure of the original clinical dataset, which comprised four primary feature groups: (1) blood pressure, (2) lipid profile, (3) renal and biochemical markers, and (4) blood cell counts. Consequently, the initial augmentation phase involved generating one IRF from each group, resulting in a total of four IRFs. This configuration was intended to explore the initial impact of structured feature expansion on model performance.

Subsequently, two additional rounds of feature augmentation were conducted by adding four more IRFs in each round, leading to three distinct experimental configurations: 4, 8, and 12 IRFs. This stepwise expansion allowed for a systematic evaluation of the effects of increasing feature dimensionality on predictive performance, model stability, and interpretability.

Although there is no established mathematical theory dictating the optimal number of augmented features, the decision to use incremental levels of IRFs was grounded in a structured exploratory design approach. This method is supported by prior studies, which suggest that progressively increasing the feature space helps identify saturation points in model performance while mitigating the risk of overfitting (Shorten & Khoshgoftaar, 2019; Jha et al., 2022; Mumuni & Mumuni, 2022).

The feature augmentation is applied to increase the features of the original clinical dataset. New features were constructed using existing equations based on medical domain knowledge, presenting the relationships between features within the same group. Therefore, these new augmented features are named Inter-Relation-based Feature or IRF. The primary objective of IRF is to enhance the explainability and trustworthiness of the proposed model by ensuring that the newly constructed features are based on clinically meaningful and interpretable relationships. The experiments involved randomly augmenting the data using 4, 8, and 12 features sequentially, to determine the most effective set for classification. The details of new features used for constructing IRF are presented in Table 3.2.

**Table 3.2** IRF augmented features

| No. | Features Detail | Description |
|---|---|---|
| 1. | Average blood pressure (ABP) | ABP, a key indicator of circulation, is calculated from systolic and diastolic pressures during heart contraction and relaxation, respectively. |
| 2. | Cholesterol-HDL Ratio (CHR) | The CHD ratio, used to assess cardiovascular risk, is calculated by dividing the total cholesterol level by the HDL level. |
| 3. | Neutrophil to Lymphocyte Ratio (NLR) | NLR, used to assess inflammation and immune response, is calculated by dividing the number of neutrophils by the number of lymphocytes. It is commonly applied in the evaluation of chronic diseases and cancer. |

**Table 3.2** (continued)

| No. | Features Detail | Description |
|---|---|---|
| 4. | Modification of Diet in Renal Disease (MDRD) | MDRD, used to assess kidney function, is calculated from serum creatinine adjusted for age, gender, and ethnicity, and is commonly used in the management of chronic kidney disease. |
| 5. | Neutrophil Count (NC) | NC, used to assess immune function, measures the number of neutrophil white blood cells in the blood. |
| 6. | Triglyceride-HDL Ratio (TG/HDL Ratio) | TG/HDL, used to assess cardiovascular risk and insulin resistance, is calculated by dividing triglyceride levels by HDL cholesterol; higher ratios indicate greater risk. |
| 7. | Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) | The CKD-EPI equation, used for accurate assessment of kidney function, improves upon the MDRD formula by incorporating serum creatinine, age, gender, and ethnicity, thereby enhancing the diagnosis of chronic kidney disease. |
| 8. | HDL-LDL ratio (HDL/LDL Ratio) | The HDL/LDL Ratio, used to assess cardiovascular health, is calculated by dividing HDL by LDL; higher ratios indicate a better cholesterol balance and a reduced risk. |
| 9 | Mean Arterial Pressure (MAP): | MAP represents the average pressure in the arteries during one cardiac cycle. It is a better indicator of blood flow to organs than systolic or diastolic pressure alone. A normal MAP ensures proper blood supply to vital organs. |
| 10 | Pulse Pressure (PP): | PP reflects the difference between systolic and diastolic blood pressure. A high PP may indicate stiff arteries, while a low PP may suggest reduced blood flow. It is a valuable marker for cardiovascular health. |

**Table 3.2** (continued)

| No. | Features Detail | Description |
|---|---|---|
| 11 | Atherogenic Index of Plasma (AIP): | AIP measures the risk of cardiovascular disease by evaluating the balance between harmful triglycerides and beneficial HDL cholesterol. A higher AIP indicates a higher risk of atherosclerosis (plaque buildup in arteries). |
| 12 | Fasting Glucose to HDL Ratio: | This ratio evaluates the relationship between fasting blood sugar (FBS) and HDL cholesterol. A higher ratio indicates a higher risk of insulin resistance and cardiovascular disease, as it reflects poor glucose metabolism and low levels of protective HDL cholesterol. |

From Table 3.2, the related equations are described as follows:

ABP (Jameson et al., 2022) is calculated by using Equation (3.1).

$$BP = \frac{SBP+D}{2} \tag{3.1}$$

HDL/LDL Ratio, which was a crucial indicator of cardiovascular disease risk, was computed by dividing total cholesterol by HDL cholesterol (Jameson et al., 2022; Skerrett, 2014; Bishop et al., 2023). The necessary data for this calculation is obtained from the lipid levels group, and the formula is given by Equation (3.2).

$$\text{Cholesterol} - \text{HDL Ratio} = \frac{Cholesterol}{HDL} \tag{3.2}$$

NLR was a ratio of neutrophils to lymphocytes in white blood cells (Jameson et al., 2022; Bishop et al., 2023; Hall, 2021). This ratio is calculated using data from the blood cell group and the formula represented by Equation (3.3).

$$NLR = \frac{\text{Neutrophil}\div 100\times\text{WBC}}{\text{Lymphocyte} \div 100\times\text{WBC}} \tag{3.3}$$

MDRD was employed to estimate glomerular filtration rate (eGFR), a measure of kidney function, especially in patients with chronic kidney disease (Jameson et al., 2022; Skerrett, 2014). The data required for this calculation is from the minerals and chemical substances group, and the formula is given by Equation (3.4).

$$MDRD = 175 \times \text{Creatinine}(-1.154) \times \text{Age}(-0.203) \tag{3.4}$$

NC was calculated using the percentage of neutrophils in a complete blood count and the total white blood cell count (Bishop et al., 2023; Hall, 2021). The data for this calculation is obtained from the blood cell group, and the formula is given by Equation (3.5).

$$\text{Neutrophil Count } = \frac{Neutrophi}{100} \times WBC \tag{3.5}$$

TG/HDL Ratio was another risk factor for cardiovascular disease and metabolic disorders (Jameson et al., 2022). It is calculated using data from a lipid level group, and the formula is given by Equation (3.6).

$$\text{Triglyceride} - \text{HDL Ratio} = \frac{\text{Triglyceride}}{HDL} \tag{3.6}$$

The CKD-EPI formula was an updated equation for estimating eGFR (Jameson et al., 2022). The data required for this calculation are from the minerals and chemical substances group, and separate formulas are provided for males and females, represented by Equation (3.7) for females and Equation (3.8) for males, respectively.

$$eGFR = 141 \times \min\left(\frac{creatinine}{0.7}, 1\right)^{-0.329} \times max\left(\frac{creatinine}{0.7}, 1\right)^{-1.209} \times (0.993)^{age} \times 1.018 \times 1.159^{if\ Black} \tag{3.7}$$

$$eGFR = 141 \times \min\left(\frac{creatinine}{0.7}, 1\right)^{-0.411} \times max\left(\frac{creatinine}{0.7}, 1\right)^{-1.209} \times (0.993)^{age} \times 1.018 \times 1.159^{if\ Black} \tag{3.8}$$

HDL/LDL Ratio is used to assess vascular health and cardiovascular risk. It is calculated by dividing HDL cholesterol by low-density lipoprotein (LDL) cholesterol. The data for this calculation is from lipid group, and the formula is given by Equation (3.9).

$$\mathrm{HDL} - \mathrm{LDL\ Ratio} = \frac{HDL}{LDL} \tag{3.9}$$

**Mean Arterial Pressure (MAP)** was an estimate of the average pressure in a person's arteries during one cardiac cycle (Hall, 2021). It is crucial in evaluating tissue perfusion. The formula is represented in Equation (3.10).

$$\mathrm{MAP} \ = \mathrm{DBP} + \frac{1}{3}(SBP - DBP) \tag{3.10}$$

**Pulse Pressure (PP)** represented the force that the heart generates with each contraction (Jameson et al., 2022). It is calculated as the difference between systolic and diastolic pressure, as shown in Equation (3.11).

$$\mathrm{PP} \ = \mathrm{SBP} - \mathrm{DBP} \qquad 3.11)$$

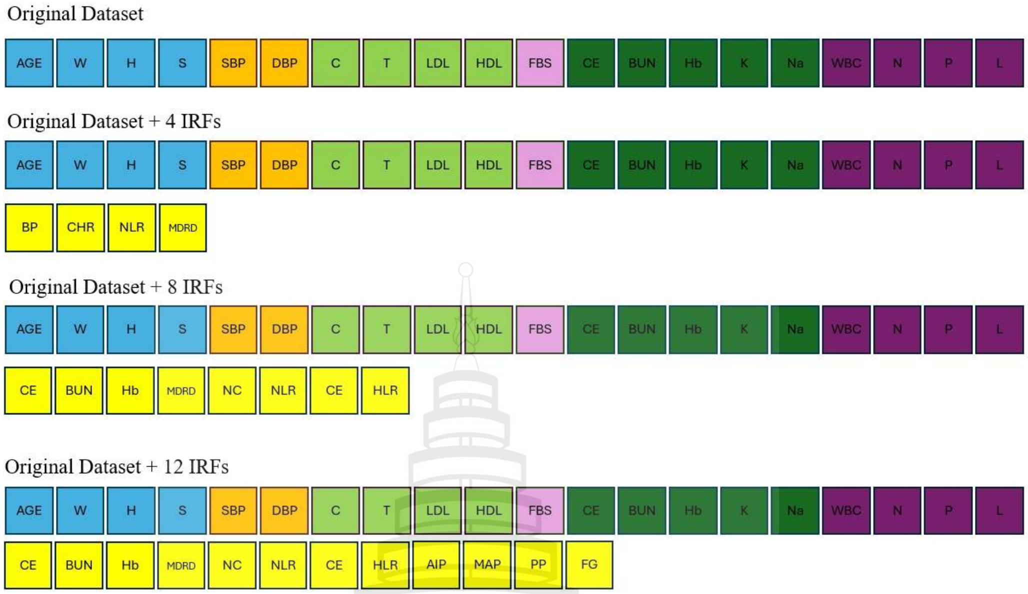
**AIP** was a logarithmic index used to assess cardiovascular risk based on lipid profile (Bishop et al., 2023). It is calculated using the triglycerides to HDL cholesterol ratio in mmol/L as shown in Equation (3.12).

$$AIP = \log_{10}\left(\frac{Triglycerides}{HDL-C}\right) \tag{3.12}$$

**The Fasting Glucose to HDL Ratio** is a simple index that may indicate insulin resistance or risk of metabolic syndrome (Hall, 2021). The formula is shown in Equation (3.13).

$$Glucose - to - HDL = \frac{Fasting\ Glucose}{HDL-C} \tag{3.13}$$

Figure 3.4 illustrates the various combination datasets used to investigate the proposed model in this study, including the original data, the original data + 4 IRFs, the original data + 8 IRFs, and the original data + 12 IRFs.


Original Dataset
AGE W H S SBP DBP C T LDL HDL FBS CE BUN Hb K Na WBC N P L
Original Dataset + 4 IRFs
AGE W H S SBP DBP C T LDL HDL FBS CE BUN Hb K Na WBC N P L
BP CHR NLR MDRD
Original Dataset + 8 IRFs
AGE W H S SBP DBP C T LDL HDL FBS CE BUN Hb K Na WBC N P L
CE BUN Hb MDRD NC NLR CE HLR
Original Dataset + 12 IRFs
AGE W H S SBP DBP C T LDL HDL FBS CE BUN Hb K Na WBC N P L
CE BUN Hb MDRD NC NLR CE HLR AIP MAP PP FG

**Source** Chaiyo et al. (2025)

**Figure 3.4** An example of datasets illustrating feature grouping and inter-relation-based feature (IRF) construction

From the conceptual diagram of the feature combination set, the example of a dataset for constructing the model, having the combination of original features and IRF, is shown in Table 3.3.

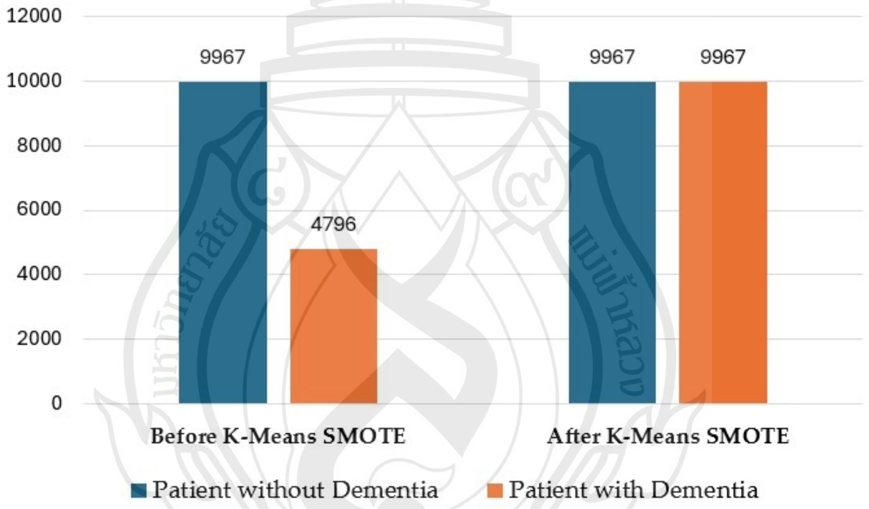
**Table 3.3** Examples of the original dataset with 4 IRFs

| Original Features | | | | | | | | | | | | | | | | | | | | IRFs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AGE** | **W** | **H** | **S** | **SBP** | **DBP** | **Chol** | **TG** | **LDL** | **HDL** | **FBS** | **Cr** | **BUN** | **Hb** | **K** | **Na** | **WBC** | **Neut** | **Plt** | **…** | **ABP** | **CHR** | **NLR** | **MDRD** |
| 74 | 60.73 | 153.02 | 0 | 128.38 | 64.84 | 145.67 | 86.94 | 89.57 | 41.7 | 116.28 | 1.03 | 17.63 | 11.078 | 4.20 | 137.09 | 12210 | 91.09 | 292530 | … | 193.22 | 0.4655 | 1.6263 | 2.0848 |
| 74 | 54.56 | 147.21 | 0 | 129.39 | 62.59 | 170.64 | 133.04 | 119.45 | 42.28 | 158.7 | 0.97 | 16.93 | 12.259 | 3.90 | 139.36 | 12276 | 91.65 | 287970 | … | 191.98 | 0.3539 | 1.4285 | 3.1466 |
| 74 | 61.34 | 155.71 | 0 | 127.37 | 63.89 | 146.03 | 81.67 | 89.81 | 40.19 | 99.48 | 1.0045 | 20.07 | 10.969 | 4.09 | 137.15 | 10725 | 91.30 | 246440 | … | 191.26 | 0.4475 | 1.6259 | 2.0320 |
| 83 | 47.16 | 143.69 | 1 | 134.49 | 58.61 | 170.93 | 120.31 | 119.3 | 43.95 | 142.23 | 1.75 | 28 | 9.400 | 2.40 | 137.80 | 10862 | 86.40 | 390000 | … | 193.10 | 0.3683 | 1.4327 | 2.7374 |

After the feature augmentation process, data standardization was conducted, and the unbalanced data was handled for the subsequent process.

### 3.2.4 Data Standardization and Balancing

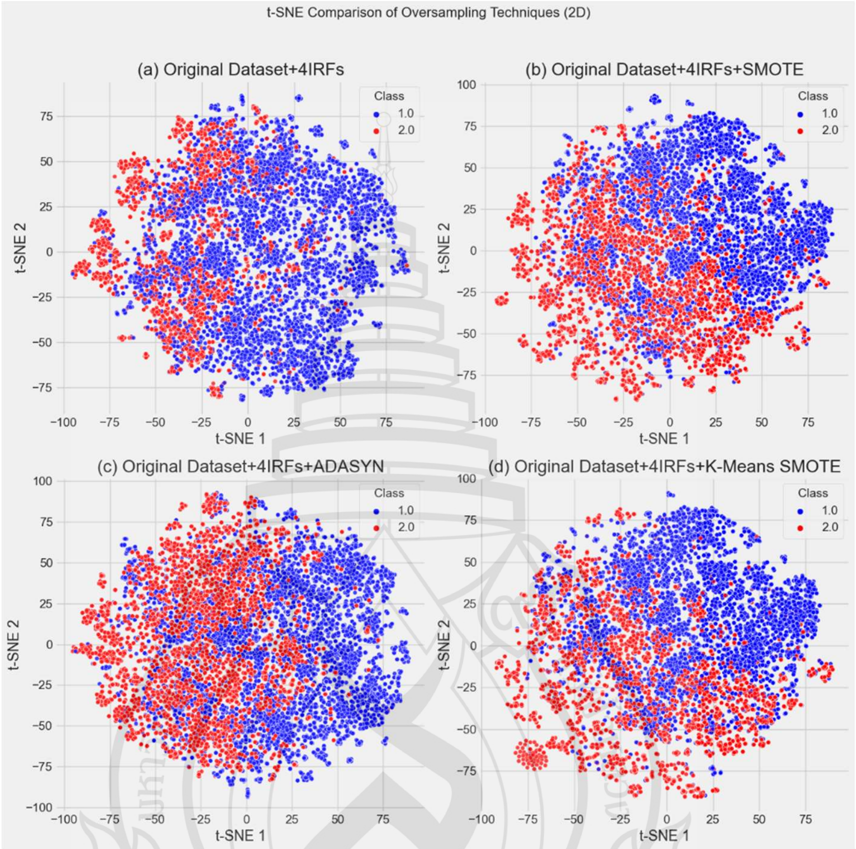
Features with different scales (e.g., SBP vs. WBC) can introduce bias in ML models. Without standardization, features with larger numerical ranges may dominate the learning process. For this study, data standardization was used to enhance the distribution of the data, ensuring a mean of zero and a standard deviation of one, thereby achieving a normal distribution of the data. In this study, K-Means SMOTE was selected over traditional SMOTE and ADASYN due to its ability to generate synthetic samples in more representative and safer feature space regions, reducing the noise risk and enhancing model generalization, particularly in imbalanced clinical datasets. In this study, the data of patients with dementia is a minority group. After applying K-Means SMOTE, the number of minority instances increases, as shown in Figure 3.5.



**Source** Chaiyo et al. (2025)

**Figure 3.5** Dataset before and after applying K-Means SMOTE. The minority class, representing patients with dementia, was resampled to match the number of instances in the majority class, representing patients without dementia

To verify structural differences in synthetic data, t-SNE was used to visualize datasets after applying IRFs with SMOTE, K-Means SMOTE, and ADASYN, as shown in Figure 3.6.



**Source** Chaiyo et al. (2025)

**Figure 3.6** t-SNE visualization of original dataset + 4IRF (a), applying SMOTE (b), applying ADASYN (c), and applying K-Means SMOTE. The blue dots are patients without dementia, and the red dots are patients with dementia

Figure 3.6(a) shows the imbalanced 4IRF dataset, with minority class samples (red) sparsely distributed, limiting effective boundary learning. In Figure 3.6(b), SMOTE improves the balance by spreading minority samples, but it introduces overlap with the majority class, which risks reduced precision. Figure 3.6(c) shows ADASYN creating irregular minority distributions near class boundaries, increasing overlap and noise. In contrast, Figure 3.6(d) illustrates K-Means SMOTE, which produces a more structured and clustered distribution, thereby enhancing minority representation while preserving class separability. Among the four scenarios, K-Means SMOTE demonstrates the most effective augmentation strategy for this study. It enhances class balance while preserving local structure and minimizing overlap between classes. Unlike SMOTE and ADASYN, which either over-smooth or overconcentrate synthetic data, K-Means SMOTE achieves a principled balance between coverage and clarity. Therefore, K-Means SMOTE stands out as the most reliable oversampling method based on visual and structural evidence from the t-SNE analysis. To justify the chosen oversampling method, statistical tests are presented in the results section that confirm K-Means SMOTE preserves data distribution, improves performance consistency, and enhances generalization.

### 3.2.5 Model Construction and Validation

The combined data was split 30% for testing and 70% for training. Ten-fold cross-valuation was the validation approach. Five models were compared with the ET model, including SVM, KNN, ANN, RF, and GB. The first model, SVM, is primarily applied for classification and regression applications. SVM is a supervised machine learning method that maximizes the margin between several classes by building a hyperplane in a high-dimensional space. With the most significant possible margin (Noble, 2006; Almasoud & Ward, 2019), SVM sought the optimal hyperplane that best divides the data into discrete classes.

The second model, KNN, was for classification and regression problems. KNN was a non-parametric, slow learning method. Based on the majority class or average value of its KNN in the feature space (Justin et al., 2013; Kramer, 2013), the method predicts the class or value of a given data point.

The third model, inspired by the human brain's neural architecture, ANN, was a computational model comprising layered, interconnected nodes that learn from input data to forecast results (Castellazzi et al., 2020).

The fourth model, RF, was an algorithm that generates a set of decision trees and combines the individual outputs to produce a final prediction. This method enhanced classification and regression performance by averaging the results of multiple trees, thereby reducing overfitting and increasing robustness (Salinas Ruíz, 2023).

The last model, (GB), used the gradient descent optimization method to minimize the loss function at each iteration (Kunapuli, 2023). It is an ensemble learning method that builds a strong predictive model by combining multiple weak models, typically decision trees. The algorithm constructs trees sequentially, each correcting the errors of its predecessor. In the proposed model, ET introduces additional randomness in the construction of the decision trees. By randomly selecting splits at each node, ET enhances computational efficiency while also mitigating overfitting (Zhang & Chen, 2020).

To ensure optimal model performance and enable fair comparisons across classifiers, hyperparameter tuning was conducted using grid search in combination with 10-fold cross-validation. This approach provides a systematic evaluation of predefined parameter combinations, facilitating the selection of configurations that yield the best average performance. Grid search was selected for its interpretability and exhaustive search strategy. At the same time, 10-fold cross-validation was utilized to provide robust and generalizable performance estimates, particularly critical in clinical datasets, which often exhibit moderate sample sizes and class imbalances.

### 3.2.6 Model Comparison

All necessary measurements were applied to evaluate the performance of the models, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP is the number of correct predictions where the actual outcome is positive. TN is the number of correct predictions where the actual outcome is negative. FP is the number of incorrect predictions where the actual outcome is negative but the model predicts a positive outcome. FN is the number of incorrect predictions where the actual outcome is positive but the model predicts a negative outcome. Finally, the classification performances of all models were compared in terms of accuracy, ROC curve, Area under the ROC, precision, recall, and F1-scores. These metrics are discussed in detail below.

Accuracy

Accuracy is one of the metrics used to evaluate the performance of a model or prediction. It indicates the proportion of total predictions that are correct compared to the total number of predictions. It is defined as shown in equation (3.10).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.10}$$

Precision

Precision is the ratio of correctly classified instances to all instances classified by the model (including both TP and FP). It emphasizes achieving the highest accuracy in instances predicted as positive (TP) while minimizing instances predicted as mispredicted (FP), as shown in equation (3.11).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.11}$$

Recall

Recall (or Sensitivity or TP Rate) is the ratio of correctly classified instances to all instances that truly exist. Recall counts the number of positive instances that were correctly predicted and selects them from all positive instances, including false negatives (FN). It is calculated as the ratio of TP to the sum of TP and FN, as shown in equation (3.12).

$$\text{Recall} = \frac{TP}{TP+} \tag{3.12}$$

F1-score

The F1-score is a metric used to evaluate the performance of classification models, particularly in scenarios involving imbalanced datasets where class distributions are uneven. It provides a comprehensive measure of model performance by considering both precision and recall, as shown in equation (3.13).

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{3.13}$$

ROC curve

The ROC curve is a graph used to assess the performance of a classification model by measuring its ability to distinguish between the positive class and the negative class at different threshold values. The following information is available: (1) True Positive Rate (Sensitivity, Recall): The ratio of correctly identified positive examples (TP) to the total actual positive examples (TP + FN). (2) False Positive Rate: The ratio of correctly identified negative examples (TN) to the total actual negative examples

(TN + FP), as shown in equations (3.14).

$$\text{True positive rate} = \frac{TP}{TP+FN} \tag{3.14}$$

Area under the ROC

In terms of general binary classification, the AUC represents the ability of a classifier to distinguish between two classes. Its value ranges from 0 to 1. In the case of 100% wrong predictions, the AUC is 0, and in the case of perfectly correct predictions, the AUC is 1.

# CHAPTER 4

# EXPERIMENTS AND RESULTS

## 4.1 Evaluation of Synthetic Data

This section presents the results from five key analyses: (1) evaluation of synthetic dataset effectiveness to confirm K-Means SMOTE as the optimal oversampling method, (2) assessment of the ET model's performance with various IRF combinations, (3) analysis of the confusion matrix to examine classification accuracy across classes, (4) ablation study to determine the contribution of each model component, and (5) sensitivity analysis to evaluate the influence of individual features on model predictions.

### 4.1.1 Data Distribution Similarity

Table 4.1 summarizes the Kolmogorov-Smirnov (K–S) test results across different oversampling methods. The K–S test assessed the similarity between the distributions of synthetic data and the original data by measuring the maximum difference (D-statistic) between their cumulative distribution functions (CDFs) in a univariate setting. A D-value greater than 0.05 indicated a statistically significant difference, implying that the synthetic data distribution deviates notably from the original data. This distributional shift might affect the representativeness and reliability of the oversampled data for subsequent analysis.

**Table 4.1** K-S test summarization

| Oversampling Method | Features with D > 0.05 | Max D-value | Most Affected Features |
|---|---|---|---|
| K-Means SMOTE | 3 | 0.4692 | Height (0.4692), ABP (0.3405), sbp (0.2024) |
| SMOTE | 4 | 0.1595 | Height (0.0801), dbp (0.1595), sbp (0.0527), ABP (0.0792) |
| ADASYN | 3 | 0.2474 | Height (0.2474), platelet (0.0531), na (0.0551) |

The K–S test results revealed that all oversampling methods introduced some distributional shifts, though to varying degrees. SMOTE demonstrated the closest alignment with the original dataset, affecting four features but with a relatively low maximum D-value of 0.1595, indicating minimal distributional deviation. K-Means SMOTE showed a higher maximum D-value (0.4692) for height, suggesting a more pronounced alteration in specific features, yet it affected only three variables overall. This might reflect a targeted modification near decision boundaries rather than broad data distortion. ADASYN, while affecting three features, presents a moderate D-value (0.2474) and alters features such as platelet and sodium levels, which might reflect less controlled synthetic sampling. Overall, SMOTE preserved the original data structure best, while K-Means SMOTE introduced more focused changes that could benefit model learning, and ADASYN presented more scattered and potentially noisier shifts.

### 4.1.2 Selection of Optimal Number of IRFs

In this section, the model performances were evaluated using different numbers of Inter-Relation-Based Features (4, 8, and 12 IRFs). The results from accuracy, F1-score, and AUC metrics guided the selection of the optimal number of IRFs for model development. Based on the observed trade-off between performance gain and model complexity, four IRFs were selected as the best balance to ensure both high predictive accuracy and model simplicity.

### 4.1.3 Model Performance Consistency

For this evaluation, all combined data (original data and 4IRF) applying three different oversampling methods (SMOTE, K-Means SMOTE, and ADASYN) were used to construct six different ML models, including SVM, KNN, ANN, RF, ET, and GB. The data was split into 70% for training and 30% for testing, and 10-fold cross-validation was conducted. A summary of the Friedman test's average ranked across performance metrics for all ML models using different oversampling techniques was shown in Table 4.2. Lower average ranked indicate better performance, and all differences are statistically significant ($p < 0.001$).

**Table 4.2** Summary of friedman test average ranks across cross-validation performance metrics

| Dataset | Best Model | Accuracy | Precision | Recall | F1-score | AUC-ROC | Avg Rank (Mean) |
|---|---|---|---|---|---|---|---|
| Original + 4IRFs | ET | 1.20 | 1.15 | 1.45 | 1.30 | 1.00 | 1.22 |
| Original + 4IRFs + ADASYN | ET | 1.10 | 1.00 | 1.10 | 1.00 | 1.00 | 1.04 |
| Original + 4IRFs + SMOTE | ET | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Original + 4IRFs + K-Means SMOTE | ET | 1.05 | 1.00 | 1.10 | 1.20 | 1.00 | 1.07 |

According to Table 4.2, the Friedman test results indicated that the SMOTE-augmented dataset was the most compatible with the ET model, achieving the lowest average rank (1.00) across all metrics. ADASYN (1.04) and K-Means SMOTE (1.07) also yielded strong performance. K-Means SMOTE achieved the highest F1-score (1.20), suggesting enhanced sensitivity to the minority class. These findings suggested that SMOTE provided the most consistent overall results. At the same time, K-Means SMOTE might offer performance gains in recall and F1-score, making it a valuable alternative when prioritizing sensitivity to minority class predictions.

### 4.1.4 Generalization Capability

The remaining 30% unseen data, created from 3 oversampling methods, was used for testing all 6 ML models. The performance metrics of the best model on unseen testing data, using different oversampling techniques, are presented in Table 4.3.

**Table 4.3** Performance of classification models on unseen test data across oversampling techniques

| Dataset | Best Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| Original | ET | 95.15% | 95.98% | 93.00% | 94.29% | 98.85% |
| Original + ADASYN | ET | 94.74% | 93.85% | 94.21% | 94.03% | 98.76% |
| Original + SMOTE | ET | 94.94% | 94.72% | 93.68% | 94.17% | 98.73% |
| Original + K-Means SMOTE | ET | 95.26% | 95.68% | 93.48% | 94.47% | 98.81% |

On the unseen testing dataset, the ET model achieved its best generalization performance when trained with K-Means SMOTE, reaching the highest Accuracy (95.26%) and F1-score (94.47%) among all datasets. While SMOTE and ADASYN also improved predictive outcomes, K-Means SMOTE's cluster-driven oversampling likely enhanced class boundary learning, resulting in superior performance on unseen data. This confirms its effectiveness as the most suitable oversampling method for optimizing ET in real-world applications.

### 4.1.5 Optimal Oversampling Method

Considering the results from the K–S test, Friedman rankings, and unseen testing performance, K-Means SMOTE was found to be the optimal oversampling method for this study. While SMOTE demonstrated the closest distributional similarity to the original data (lowest D-values in the K–S test) and achieved the best average ranking in cross-validation (Friedman test), K-Means SMOTE provided the highest generalization performance on unseen data, with the best accuracy (95.26%) and F1-score (94.47%). The results confirmed that the clustering-targeting approach creates more informative synthetic samples near decision boundaries, thereby enhancing class representation without introducing excessive noise. K-Means SMOTE offers the most efficient trade-off between improving model discriminability and maintaining distributional integrity, making it the best option for this study. The K-Means SMOTE method was therefore selected for further investigation about the effects of IRFs, aiming to achieve optimal model performance.

## 4.2 Evaluation of Effective Classification Model

Following the identification of K-Means SMOTE as the most effective oversampling method in this study, an additional investigation was conducted to determine the optimal combination of model and dataset. Specifically, this analysis aimed to evaluate which input feature set comprising either 4 IRFs, 8 IRFs, or 12 IRFs yields the best performance across various machine learning models.

### 4.2.1 Descriptive Summary of 10-Fold Cross-Validation Results

Table 4.4 - 4.6 shows the descriptive representation of the 10-fold cross-validation of all ML models applying the original dataset, the original set with 4IRF applying K-Means SMOTE, the original dataset with 8IRF applying K-Means SMOTE, and the original dataset with 12IRF applying K-Means SMOTE, respectively.

**Table 4.4** Model performance based on the stratified 10-fold cross-validation on the original dataset

| **Model** | **Original Dataset Feature (n = 20)** | | | | |
|---|---|---|---|---|---|
| | **Precision (%)** | **Recall (%)** | **F1 Score (%)** | **Accuracy (%)** | **AUC** |
| GB | 0.93 ± 0.01 | 0.92 ± 0.01 | 0.93 ± 0.01 | 0.94 ± 0.01 | 0.98 ± 0.00 |
| RF | 0.95 ± 0.01 | 0.92 ± 0.01 | 0.94 ± 0.01 | 0.95 ± 0.01 | 0.99 ± 0.00 |
| ET | 0.96 ± 0.01 | 0.93 ± 0.01 | 0.94 ± 0.01 | 0.95 ± 0.01 | 0.99 ± 0.00 |
| SVM | 0.34 ± 0.00 | 0.50 ± 0.00 | 0.40 ± 0.00 | 0.68 ± 0.00 | 0.78 ± 0.01 |
| KNN | 0.86 ± 0.01 | 0.84 ± 0.01 | 0.85 ± 0.01 | 0.87 ± 0.01 | 0.92 ± 0.01 |
| ANN | 0.55 ± 0.14 | 0.58 ± 0.09 | 0.53 ± 0.12 | 0.68 ± 0.02 | 0.76 ± 0.02 |

**Table 4.5** Model performance based on the stratified 10-fold cross-validation on the original dataset with 4IRF and K-Means SMOTE

| Model | Original Dataset with 4IRFs and K-Means SMOTE (n = 24) | | | | |
|---|---|---|---|---|---|
| | Precision (%) | Recall (%) | F1 Score (%) | Accuracy (%) | AUC |
| GB | 0.95 ± 0.01 | 0.95 ± 0.01 | 0.95 ± 0.01 | 0.95 ± 0.01 | 0.99 ± 0.00 |
| RF | 0.96 ± 0.00 | 0.96 ± 0.00 | 0.96 ± 0.00 | 0.96 ± 0.00 | 0.99 ± 0.00 |
| ET | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.99 ± 0.00 |
| SVM | 0.76 ± 0.01 | 0.75 ± 0.01 | 0.74 ± 0.01 | 0.75 ± 0.01 | 0.83 ± 0.01 |
| KNN | 0.90 ± 0.01 | 0.90 ± 0.01 | 0.90 ± 0.01 | 0.90 ± 0.01 | 0.96 ± 0.01 |
| ANN | 0.75 ± 0.05 | 0.68 ± 0.08 | 0.64 ± 0.13 | 0.68 ± 0.08 | 0.77 ± 0.07 |

**Table 4.6** Model performance based on the stratified 10-fold cross-validation on the original dataset with 8IRF and K-Means SMOTE

| Model | Original Dataset with 8IRFs and K-Means SMOTE (n = 28) | | | | |
|---|---|---|---|---|---|
| | Precision %) | Recall (%) | F1 Score (%) | Accuracy (%) | AUC |
| GB | 0.95 ± 0.00 | 0.95 ± 0.00 | 0.95 ± 0.00 | 0.95 ± 0.00 | 0.99 ± 0.00 |
| RF | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.99 ± 0.00 |
| ET | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.96 ± 0.01 | 0.99 ± 0.00 |
| SVM | 0.77 ± 0.01 | 0.76 ± 0.01 | 0.75 ± 0.01 | 0.76 ± 0.01 | 0.84 ± 0.01 |
| KNN | 0.91 ± 0.01 | 0.91 ± 0.01 | 0.91 ± 0.01 | 0.91 ± 0.01 | 0.96 ± 0.01 |
| ANN | 0.79 ± 0.04 | 0.76 ± 0.06 | 0.75 ± 0.06 | 0.76 ± 0.06 | 0.82 ± 0.05 |

The comparison of model performance across the three tables reveals that incorporating additional IRF (Imbalance Ratio Handling) features with K-Means SMOTE generally improves classification metrics, with 4IRF (Table 4.5) and 8IRF (Table 4.6) both outperforming the original dataset (Table 4.4). Tree-based models (GB, RF, ET) already performed well initially but saw marginal gains in F1-score and AUC (e.g., RF improved from 0.94 to 0.96 F1). However, the most notable improvements were observed in weaker models, such as SVM (F1 score increased from 0.40 to 0.75) and ANN (F1 score rose from 0.53 to 0.75 with 8IRF), suggesting that

IRF features enhance robustness for less stable algorithms. While 4IRF and 8IRF yielded similar results for most models, 8IRF provided slight advantages for KNN (F1: 0.91 vs. 0.90) and ANN (F1: 0.75 vs. 0.64), indicating that higher IRF counts might further stabilize performance, particularly for non-tree models. In terms of ensemble learning models, 4IRF is sufficient, as it achieves near-peak performance without unnecessary feature expansion.

### 4.2.2 Model Performance Based on Accuracy

The testing results on unseen data in terms of accuracy are illustrated in Table 4.7.

**Table 4.7** Accuracy comparison

| Accuracy (%) | | | | | |
|---|---|---|---|---|---|
| **Model** | **Original Dataset** | **Original Dataset + K-means SMOTE** | **Original Dataset + K-means SMOTE + 4 IRF** | **Original Dataset + K-means SMOTE + 8 IRF** | **Original Dataset + K-means SMOTE + 12 IRF** |
| SVM | 93.05 | 94.73 | 94.52 | 94.50 | 94.55 |
| GB | 93.84 | 94.58 | 94.67 | 95.10 | 94.77 |
| ET | 95.08 | 96.39 | 96.47 | 96.52 | 96.51 |
| RF | 94.63 | 96.02 | 96.10 | 96.27 | 96.00 |
| KNN | 91.24 | 92.58 | 92.44 | 92.56 | 91.99 |
| ANN | 92.68 | 94.70 | 94.95 | 94.90 | 95.17 |

Table 4.7 illustrates the classification performance of each model in terms of accuracy. The results showed that the combined dataset (applying K-Means SMOTE and adding IRF) raised the accuracy of all models compared to the original dataset. The employment of K-Means SMOTE alone showed a clear boost in accuracy across all models, therefore proving the value of extra pertinent features.  Most models continue to show a minor increase as more features (4 IRF and 8 IRF) are added. GB, ET, and RF benefit most and achieve their best accuracy at 8 IRF. Being ensemble-based models, ET and RF exhibit the highest accuracy gains, indicating that they effectively utilize the recently acquired characteristics to enhance classification. Conversely, SVM, KNN, and ANN achieve very modest gains after applying K-Means SMOTE,

suggesting that these models may not fully utilize additional information beyond a certain point. With ensemble models providing the most significant improvement, K-Means SMOTE, and IRF generally enhance accuracy; however, other models exhibit declining returns as more features are included.

Notably, when 12 IRFs are introduced (*Original Dataset + K-means SMOTE + 12 IRFs),* only ANN and KNN continue to show clear improvement, achieving their highest accuracies of 95.17% and 91.99%, respectively. In contrast, ensemble models like RF, ET, and GB show a slight decline or plateau, suggesting potential overfitting or redundancy beyond 8 IRF. This suggests that while IRFs are beneficial, the optimal number may vary depending on the model type and complexity.

### 4.2.3 Model Performance Based on Precision

The testing results on unseen data in terms of precision are shown in Table 4.8.

**Table 4.8** Precision comparison

| | Precision (%) | | | | |
|---|---|---|---|---|---|
| **Model** | **Original Dataset** | **Original Dataset + K-means SMOTE** | **Original Dataset + K-means SMOTE + 4 IRF** | **Original Dataset + K-means SMOTE + 8 IRF** | **Original Dataset + K-means SMOTE + 12 IRF** |
| SVM | 94.35 | 94.26 | 94.25 | 94.05 | 94.11 |
| GB | 94.44 | 93.98 | 94.07 | 93.92 | 93.98 |
| ET | 94.01 | 94.72 | 94.79 | 94.47 | 94.42 |
| RF | 93.66 | 94.14 | 93.99 | 93.82 | 93.95 |
| KNN | 97.34 | 97.55 | 97.46 | 97.57 | 97.59 |

Table 4.8 displays the classification performance of each model in terms of precision. Depending on the model, the combined dataset affected precision differently; some models gained from the extra characteristics, while others experienced minor declines or swings. As more features were introduced, KNN and ANN exhibited a continuous increase in accuracy, implying that these models could efficiently utilize the additional information to enhance classification performance. With K-Means SMOTE and 4 IRF, et al. showed an initial increase in precision; however, at K-Means SMOTE and 8 IRF, there is a slight decline, suggesting that

adding too many features might generate noise rather than improve precision. Conversely, SVM, GB, and RF exhibit modest decreases in precision following the inclusion of K-Means SMOTE and extra IRF, suggesting that for these models, the increased features may slightly increase false positives, resulting in a minor drop in precision.

Interestingly, when 12 IRFs are included, KNN and ANN continue to improve, reaching their highest precision of 97.59% and 94.96%, respectively, confirming their ability to benefit from expanded feature sets. Meanwhile, SVM, GB, ET, and RF either remain stable or show slight declines, suggesting that their optimal precision might have already been reached at earlier stages (e.g., 4 or 8 IRF), beyond which the added features offer diminishing or even adverse effects on precision.

### 4.2.4 Model performance based on Recall

The testing results on unseen data in terms of recall are shown in Table 4.9.

**Table 4.9** Recall comparison

| | Recall (%) | | | | |
|---|---|---|---|---|---|
| **Model** | **Original Dataset** | **Original Dataset + K-means SMOTE** | **Original Dataset + K-means SMOTE + 4 IRF** | **Original Dataset + K-means SMOTE + 8 IRF** | **Original Dataset + K-means SMOTE + 12 IRF** |
| SVM | 94.96 | 94.22 | 93.85 | 94.16 | 94.02 |
| GB | 96.53 | 95.31 | 95.11 | 95.49 | 95.57 |
| ET | 98.79 | 97.97 | 97.86 | 97.57 | 97.77 |
| RF | 98.55 | 97.76 | 97.66 | 97.34 | 97.58 |
| KNN | 93.40 | 91.03 | 90.94 | 91.13 | 91.43 |
| ANN | 94.85 | 94.35 | 94.34 | 94.36 | 94.34 |

Table 4.9 displays the classification performance of each model in terms of recall. It can be observed that different models have varying effects on recall from the combined dataset; some exhibit a drop, while others remain relatively stable. After the inclusion of K-Means, SMOTE, SVM, GB, ET, RF, and KNN, all exhibit a drop in recall. A further decrease in performance as IRF features were added indicates that these models might suffer from overfitting or noise generated by the extra features,

hence producing false negatives. Initially, ET and RF showed the most significant decreases when K-Means SMOTE was applied and IRF was added. This result suggested that the feature expansion somewhat reduces their capacity to label positive events correctly. Notably, KNN showed a consistent decline, likely due to its sensitivity to feature dimensionality. Conversely, ANN remained relatively constant and exhibited only slight variations, suggesting that neural networks may be more resilient to feature expansion in terms of recall. In terms of most models, the combined dataset reduced overall recall, suggesting that even if extra features might increase other measures, such as accuracy or precision, they could also introduce complexity that made it more difficult for models to capture positive examples adequately.

When 12 IRFs are added (Original Dataset + K-means SMOTE + 12 IRFs), most models exhibit stabilization or a slight improvement in recall. ET improved slightly to 97.77%, regaining some performance after the earlier drop, while RF also recovered to 97.58%. KNN, despite its previous consistent decline, increased modestly to 91.43%. On the other hand, SVM continued to drop slightly to 94.02%, and ANN remained stable at 94.34%. These results indicated that while additional features might introduce noise for some models, others could adapt and benefit marginally from extended feature sets at this stage.

Overall, ET and RF perform best at 4–8 IRFs, showing high recall with minimal loss before plateauing or slightly recovering at 12 IRFs. ANN demonstrated consistent robustness across all feature levels, making it suitable even up to 12 IRFs. In contrast, models like SVM and GB tended to perform best at 4 IRFs and decline thereafter, while KNN might benefit from 12 IRFs after earlier reductions. However, its performance remained sensitive to feature expansion. This suggested model-specific thresholds for optimal feature inclusion, balancing information gain and noise.

### 4.2.5 Model Performance Based on F1 Score

The testing results on unseen data in terms of F1-score are shown in Table 4.10.

**Table 4.10** Comparison with F1 score

| F1 Score (%) | | | | | |
|---|---|---|---|---|---|
| **Model** | **Original Dataset** | **Original Dataset + K-means SMOTE** | **Original Dataset + K-means SMOTE + 4 IRF** | **Original Dataset + K-means SMOTE + 8 IRF** | **Original Dataset + K-means SMOTE + 12 IRF** |
| SVM | 94.66 | 94.24 | 94.05 | 94.11 | 94.07 |
| GB | 95.47 | 94.64 | 94.59 | 94.70 | 93.98 |
| ET | 96.34 | 96.32 | 96.30 | 96.00 | 96.07 |
| RF | 96.04 | 95.92 | 95.79 | 95.55 | 95.73 |
| KNN | 95.33 | 94.18 | 94.09 | 94.24 | 94.41 |
| ANN | 94.81 | 94.69 | 94.72 | 94.75 | 94.65 |

Table 4.10 displays the classification performance of each model in terms of F1-score. The combined dataset affected the F1-score differently depending on the model; most models only caused modest changes or slight reductions. After the addition of IRF and the application of K-Means SMOTE, SVM, GB, ET, RF, and KNN all showed a drop in F1-score, indicating that the trade-off between precision and recall was uneven, most likely due to a loss in recall. KNN and RF showed the most significant declines, implying that these models might have struggled with the additional features, possibly due to overfitting or an increased number of false negatives. ET maintains a relatively constant F1-score, indicating that it is robust against feature expansion. GB exhibits a minor decline but somewhat recovers at the combined dataset of K-Means SMOTE and 8 IRF. With just minor variations, ANN remained almost constant, demonstrating that neural networks could better manage the extra features than other models. Although some models adapted better than others, the overall combined dataset has a somewhat negative impact on the F1-score. This implied that although feature development could enhance accuracy and precision, it might not continually improve the balance between precision and recall, which was essential for maintaining a high F1-score.

When 12 IRFs were included (Original Dataset + K-means SMOTE + 12 IRF), most models exhibited stabilization or minor improvement in F1-score. ET reached its highest score at this point (96.07%), confirming its resilience to extended features. ANN also maintained a strong and stable F1-score (94.65%), reflecting its robustness across feature levels. In contrast, GB and SVM declined slightly to 93.98% and 94.07% respectively, suggesting diminished returns or overfitting. RF and KNN slightly recovered to 95.73% and 94.41% from their lowest points at 8 IRF.

Overall, ET benefits the most from all levels of IRF and performs best at 12 IRFs. ANN remains consistently strong across 4 to 12 IRFs, with minimal fluctuation. RF achieves its optimal F1-score at 4 IRFs (95.79%) before declining. KNN performs best at 4 IRFs but remains acceptable through 12. SVM and GB show their highest scores with the original or 4 IRF datasets, and decline thereafter. These findings emphasized that ensemble models, such as ET, could scale with added features, while models like SVM and GB might require careful feature selection to maintain F1 performance.

### 4.2.6 Model Performance Based on Average AUC-ROC

The testing results on unseen data in terms of accuracy are shown in Table 4.11.

**Table 4.11** Average AUC-ROC comparison

| | Average AUC-ROC (%) | | | | |
|---|---|---|---|---|---|
| **Model** | **Original Dataset** | **Original Dataset + K-means SMOTE** | **Original Dataset + K-means SMOTE + 4 IRF** | **Original Dataset + K-means SMOTE + 8 IRF** | **Original Dataset + K-means SMOTE + 12IRF (%)** |
| SVM | 97.10 | 98.59 | 98.57 | 98.68 | 98.61 |
| GB | 98.03 | 98.87 | 98.89 | 98.91 | 98.87 |
| ET | 98.74 | 99.49 | 99.51 | 99.45 | 99.43 |
| RF | 98.59 | 99.39 | 99.35 | 99.33 | 99.33 |
| KNN | 96.20 | 97.12 | 96.82 | 96.96 | 96.97 |
| ANN | 96.98 | 98.44 | 98.56 | 98.44 | 98.52 |

Table 4.11 displays the classification performance of each model in terms of AUC-ROC. With most models demonstrating an improvement over the original dataset, the AUC-ROC values suggest that generally adding IRF and applying K-Means

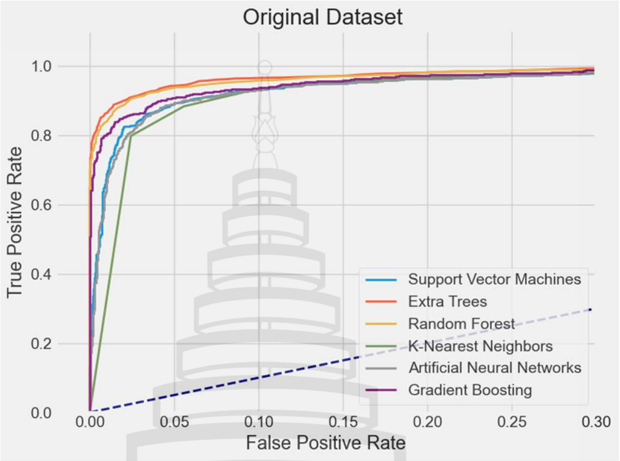
SMOTE to the original dataset improves the models' capacity to differentiate between classes. Adding K-Means SMOTE causes SVM, GB, ET, RF, and ANN to all show a notable rise in AUC-ROC, indicating improved separability between classes. Being ensemble-based models, ET and RF achieve the best AUC-ROC values, peaking around 99.5%, indicating that these models effectively utilize the extra features to enhance discrimination. At the original dataset with K-Means SMOTE and 8 IRF, however, modest variations or oscillations were noted, especially in ET and RF, which would suggest that adding too many characteristics generated noise rather than helpful information. Although KNN improved with K-Means SMOTE, its performance might be sensitive to high-dimensional data, as it showed a slight decrease after 4 IRF. Conversely, ANN consistently remained high, demonstrating its capacity to utilize the additional features efficiently. Although excessive feature addition might lead to declining returns for some models, overall, K-Means, SMOTE, and IRF enhance the AUC-ROC by increasing the discriminating power of most models.

When 12 IRFs are added (Original Dataset + K-means SMOTE + 12 IRFs), most models continue to maintain high AUC-ROC performance or experience slight improvements. ET and RF sustain top scores at 99.43% and 99.33% respectively, confirming their robustness in distinguishing between classes even with extended features. ANN also rose to 98.52%, indicating consistent benefit from IRF expansion. KNN improved steadily to 97.97%, marking its best performance at 12 IRFs despite its earlier drop. However, SVM and GB showed slight declines to 98.61% and 98.57%, suggesting that these models might not gain further from additional features beyond 8 IRFs.

Overall, ET and RF performed optimally at 4–8 IRFs and continued to hold top AUC-ROC values at 12 IRFs, reflecting their strength in managing large feature sets. ANN was highly stable and showed progressive improvement across 4 to 12 IRFs. KNN performed best at 12 IRFs, showing notable improvement with feature expansion. In contrast, SVM and GB reached peak performance at 8 IRFs and experienced slight deterioration thereafter. These results suggested that ensemble models, such as ET and RF, were the most reliable across all IRF levels, while ANN and KNN could still leverage larger feature sets. Models like SVM and GB, however, required more selective feature inclusion to sustain optimal discriminative power.

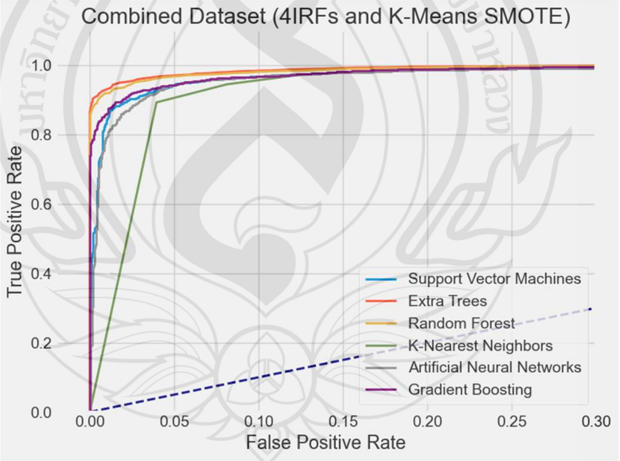
### 4.2.7 ROC Curve Comparison

Figure 4.1 – 4.4 shows the ROC curves of all models applying three different types of datasets, including the original dataset, the dataset with 4 RFs applying K-Means SMOTE, and the dataset with 8 IRFs applying K-Means SMOTE.



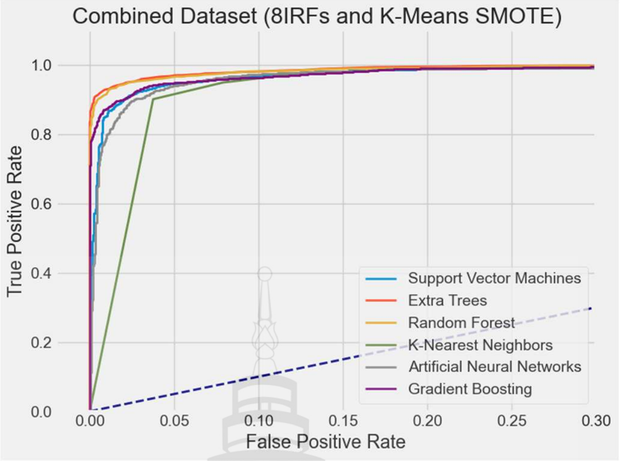
**Source** Chaiyo et al. (2025)

**Figure 4.1**. ROC curves of all models using the original dataset



**Source** Chaiyo et al. (2025)

**Figure 4.2** ROC curves of all models using the dataset with 4 IRFs and K-Means SMOTE



**Source** Chaiyo et al. (2025)

**Figure 4.3** ROC curves of all models using the dataset with 8 IRFs and K-Means SMOTE

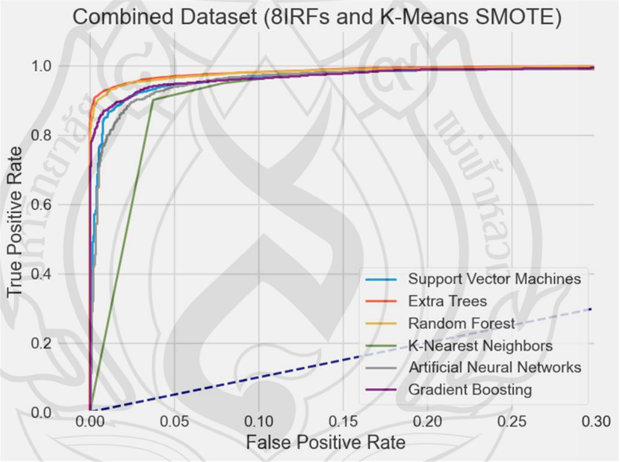


**Figure 4.4** ROC curves of all models using the dataset with 12 IRFs and K-Means SMOTE

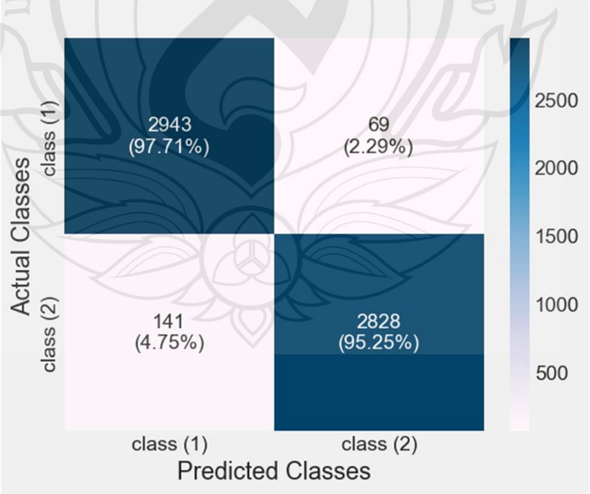
The ROC curve analysis across the original dataset and those augmented with 4 and 8 IRFs combined with K-Means SMOTE demonstrated a clear performance improvement from oversampling and feature augmentation. Ensemble models particularly ET consistently showed superior sensitivity and specificity. The dataset with 4 IRFs and K-Means SMOTE yielded the most balanced improvement, significantly enhancing the performance of non-ensemble models, such as SVM and

ANN. Although the 8 IRF configuration offered slightly higher performance, the marginal gain suggested diminishing returns, indicating that 4 IRFs provided an optimal trade-off between model complexity and effectiveness. Combining 4 IRFs with K-Means SMOTE yielded the most effective enhancement in class separability and generalization performance.

When the dataset includes 12 IRFs with K-Means SMOTE, most models retain or slightly enhance their ROC performance. ET and RF maintained their leading positions with near-perfect separability, indicating their robustness even in high-dimensional space. ANN continued to show stable and high performance, proving its resilience to increased features. KNN, despite prior sensitivity, achieved its best AUC at 12 IRFs, suggesting effective adaptation with larger feature sets. In contrast, SVM and GB exhibited marginal declines compared to their performance at 8 IRFs, which may indicate potential overfitting or a reduction in discriminative power beyond that point.

## 4.3 Confusion Matrix

Since the ET Model with 4IRF and K-Means SMOTE provided the best model for this study, its confusion matrix is shown in Figure 4.5.



**Source** Chaiyo et al. (2025)

**Figure 4.5** Performance of the extra trees model with 4IRF and K-Means SMOTE based on confusion

Figure 4.5 shows that the ET model with 4 IRFs and K-Means SMOTE accurately classified 97.91% of class 1 and 95.15% of class 2, demonstrating strong sensitivity and specificity. The low false positive and false negative rates confirmed the model's effectiveness in handling class imbalance while maintaining high discriminative power, supporting its reliability for real-world applications.

## 4.4 Ablation Study

In this study, an ablation study was conducted to assess the contribution of inter-relation-based features (IRFs), K-Means SMOTE, and ET model, as shown in Table 4.12. The ablation study involved systematically removing or modifying individual components to evaluate their impact on model performance. This approach helped confirm the additive value of each element and ensured that their inclusion meaningfully contributed to classification accuracy and clinical relevance.

**Table 4.12** Results of the ablation study

| Model Configuration | IRFs | K-means SMOTE | Classifier | Accuracy (%) | Recall (%) | Precision (%) | F1-Score (%) | AUC - ROC (%) |
|---|---|---|---|---|---|---|---|---|
| Full Model | ✓ | ✓ | ET | 96.47 | 97.86 | 94.79 | 96.30 | 99.51 |
| w/o IRFs | × | ✓ | ET | 96.30 | 98.01 | 94.73 | 96.26 | 99.49 |
| w/o SMOTE | ✓ | × | ET | 94.85 | 99.05 | 93.64 | 96.22 | 98.88 |
| w/o IRFs & SMOTE | × | × | ET | 95.08 | 98.63 | 93.96 | 96.34 | 98.74 |
| RF instead of ET | ✓ | ✓ | RF | 96.10 | 97.66 | 93.99 | 95.79 | 99.35 |

The ablation study confirmed that the whole model combining IRFs, K-Means SMOTE, and ET achieves the best performance (Accuracy: 96.47%, AUC: 99.51%). Removing IRFs led to only a slight drop, indicating a modest but supportive role, but excluding K-Means SMOTE caused a more notable decline in accuracy and AUC, highlighting its critical contribution to generalization. Even without both IRFs and SMOTE, the model maintained a high F1-score but showed reduced discriminability.

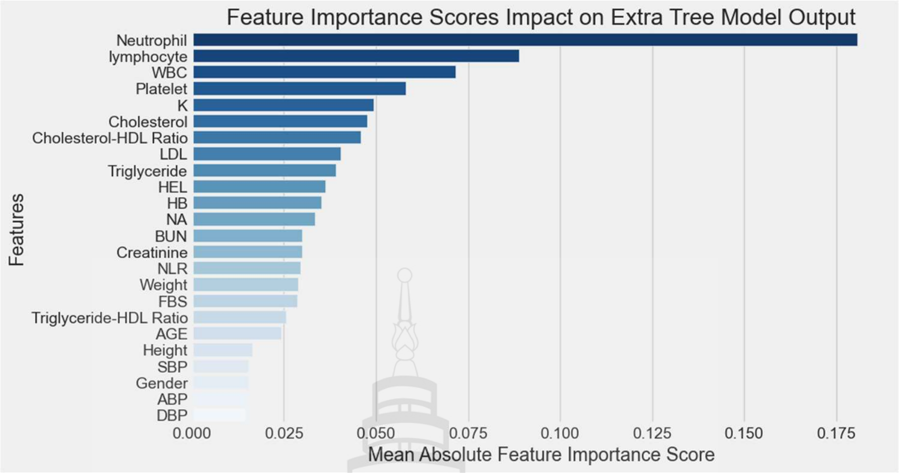
Replacing ET with RF (the second top model) slightly decreased all metrics, reaffirming ET as the most robust classifier in this setting. This ablation confirmed that while each component contributed to overall performance, K-Means SMOTE and the ET classifier were the most influential, and the addition of IRFs provided incremental improvement. The complete configuration, including IRFs, K-Means SMOTE, and Extra Trees, is validated as the optimal choice for this study.

## 4.5 Sensitivity Analysis

In this research, mean absolute feature importance scores, derived from impurity reduction in the ET model, were used as a form of global sensitivity analysis to quantify the average influence of each feature across the entire dataset. In parallel, SHapley Additive exPlanations (SHAP) values were applied to provide both global and local interpretability, enabling assessment of each feature's impact on individual predictions and supporting the evaluation of feature relevance, consistency, and model stability.

### 4.5.1 Mean Absolute Feature Importance Scores

Mean absolute feature importance scores represent a global measure of feature influence in tree-based models, indicating how frequently and effectively a feature contributes to decision splits. Figure 4.6 shows the mean absolute feature importance scores of the proposed ET model.



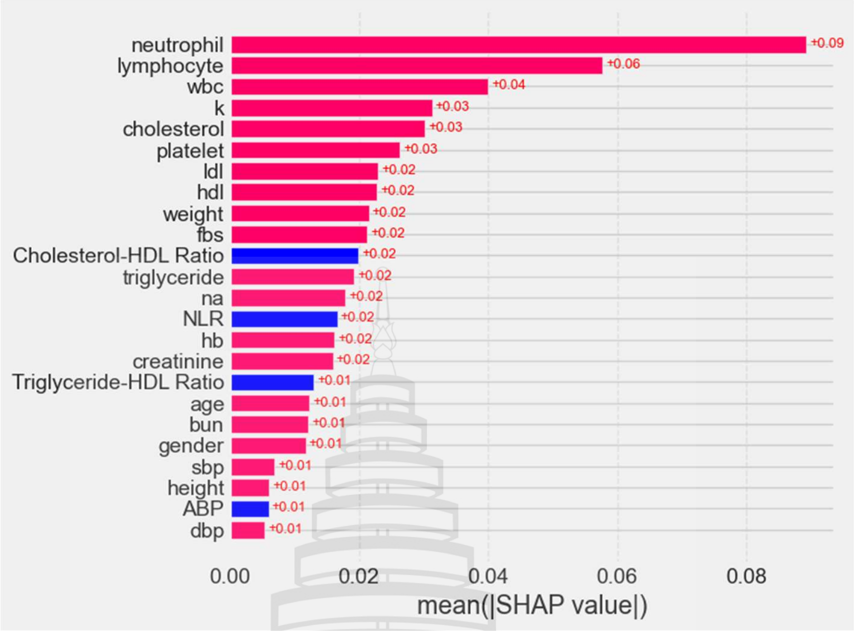
**Source** Chaiyo et al. (2025)

**Figure 4.6** Mean absolute feature importance scores of the extra trees model trained on the combined dataset (4 IRFs + K-Means SMOTE)

From Figure 4.6 shows that while Neutrophil, Lymphocyte, and WBC are the most influential features, the four IRFs (ABP, NLR, Cholesterol-HDL Ratio, and Triglyceride-HDL Ratio) also contribute to moderate importance, particularly the Cholesterol-HDL Ratio (0.0458). These findings suggested that the IRF features provided meaningful supplementary value, enhancing the model’s predictive capacity alongside key raw biomarkers.

### 4.5.2 Mean SHAP Value

At the same time, SHAP values were used for local and global sensitivity analysis in this study. It provided a model-agnostic, consistent, and interpretable sensitivity analysis based on game theory.



**Source** Chaiyo et al. (2025)

**Figure 4.7** SHAP summary plot illustrating the global and local impact of each feature on the extra trees model's predictions using the combined dataset (4 IRFs + K-Means SMOTE)

From Figure 4.7 shows that while raw biomarkers, such as neutrophils and lymphocytes, dominate predictive power, the 4 IRF features (NLR, Cholesterol-HDL Ratio, Triglyceride-HDL Ratio, and ABP) also contributed meaningfully to the model's sensitivity, with SHAP values ranging from +0.01 to +0.02. Despite their modest individual impact, these features enhanced clinical interpretability and aligned with established evidence in dementia prediction, supporting their complementary role in a model that balanced performance with clinical relevance. Elevated NLR was associated with systemic inflammation, while high CHR reflected poor vascular health, both of which are established risk factors for dementia (Zhang & Chen, 2020; Lee & Kim, 2021). These findings supported the integration of IRFs to improve the model's clinical relevance without compromising predictive performance.

In terms of model sensitivity, the analysis of Figure 4.6 and Figure 4.7 confirmed that while raw biomarkers such as neutrophil and lymphocyte exerted the most decisive influence on predictions, the 4 IRFs (ABP, NLR, Cholesterol-HDL Ratio, and Triglyceride-HDL Ratio) also contributed to the model’s sensitivity, albeit to a lesser extent. Their moderate SHAP values and importance scores indicated that slight variations in these features could still influence model output, underscoring their role as meaningful supplementary predictors. This emphasized the model’s ability to integrate both raw and composite features.

## 4.6 Discussion on the Number of Inter-Relation-Based Features (IRFs)

In this study, a total of four Inter-Relation-Based Features (IRFs) were selected for integration into the final classification model. The decision was based on a stepwise feature augmentation process, where the model's performance was evaluated using 4, 8, and 12 IRFs. As shown in Table 4.10, the use of four IRFs resulted in the best overall classification metrics, including accuracy, F1-score, and AUC. Increasing the number of IRFs beyond four led to marginal or even decreased performance, indicating that four IRFs captured the most relevant information without adding noise or redundancy.

The choice of four IRFs reflected an optimal balance between predictive performance and model simplicity. While adding more features might theoretically capture additional variability, it also increases the risk of overfitting, reducing interpretability, and introducing unnecessary model complexity. By selecting four IRFs, the study achieved stable model performance while preserving clinical interpretability, thereby facilitating practical application.

The finding was consistent with previous literature suggesting that moderate feature expansion guided by domain expertise could enhance model generalization and maintain explainability (Shorten & Khoshgoftaar, 2019; Mumuni & Mumuni, 2022). The four IRFs used in this study represented key inter-variable relationships derived from core clinical domains, contributing to a model that is both accurate and clinically relevant.

# CHAPTER 5

# DISCUSSION AND CONCLUSION

## 5.1 K-Means SMOTE Effect

Depending on their sensitivity to synthetic data and their capacity to extend from interpolated samples, K-Means SMOTE affects all models differently. Because they are robust to noise and can efficiently detect trends even in the presence of created minority class samples, ensemble models like ET and RF benefit most. Although GB also improved, it is somewhat sensitive to noisy synthetic data, which causes modest variance in accuracy. Based on margin optimization, SVM has mixed effects, as synthetic samples may introduce overlapping class borders, thereby somewhat compromising accuracy and recall. KNN suffers the most since it relies on distance-based categorization, and synthetic data can skew neighborhood ties, thereby producing either higher false positives or negatives. Although usually flexible, ANNs do not demonstrate significant increases, most likely due to inadequate hyperparameter tuning that prevents them from effectively utilizing the newly produced data. By resolving class imbalance, K-Means SMOTE increases recall and AUC-ROC for most models overall. However, its efficacy depends on how well a model can handle synthetic data without overfitting or losing precision. In conclusion, K-Means SMOTE is the most suitable oversampling method for this study, offering the best trade-off between distributional integrity and model discriminability. It achieved the highest generalization performance with the ET model.

## 5.2 IRF Effect

The ability of all models to manage more dimensionality and discover pertinent patterns from newly acquired information determines how IRF affects them. Feature augmentation helps ensemble models, such as RF and ET, the most since they can efficiently utilize extra features while maintaining stability and avoiding overfitting. Although it is more sensitive to feature noise, which can lead to variations in precision, GB also exhibits modest increases in performance. Since SVM relies on determining the optimal hyperplane in a fixed-dimensional space, adding extra features may increase complexity without enhancing class separability. Therefore, feature augmentation may not significantly benefit it. KNN suffers significantly with augmented features since it is susceptible to dimensionality, and a larger feature space may dilute strong distance correlations, thereby affecting performance. Although they can manage high-dimensional data, ANNs demonstrate no appreciable improvement, most likely because more advanced tuning is needed to derive relevant patterns from the augmented features. Overall, feature augmentation using IRFs improves performance for tree-based ensemble models but offers limited benefit for distance-based or margin-based models. Sensitivity analysis indicates that the 4 IRFs have moderate importance, serving as complementary features that support predictions alongside key biomarkers.

## 5.3 IRF and K-Means SMOTE Effect

On all models, the combined effect of K-Means SMOTE and IRF relies on their capacity to manage synthetic data and higher dimensionality. Tree-based ensemble models, such as RF and ET, yield the most significant improvements since they are designed to learn from both synthetic minority samples and additional features, thereby producing greater accuracy, recall, and AUC-ROC. Though it is somewhat subject to noise from synthetic data and feature augmentation, GB also demonstrates benefits and causes minor variations in recall and precision. SVM produces mixed results since extra features may not always help to improve hyperplane separation, and K-Means SMOTE can introduce class boundary overlaps. KNN suffers the most since both synthetic data and high-dimensional feature spaces distort distance relationships, thereby lowering its performance. Although they are usually adaptable, ANNs do not demonstrate significant gains, most likely because thorough hyperparameter tweaking is necessary to use synthetic data and new characteristics fully. While distance-based and margin-based models struggle to utilize additional data efficiently, K-Means SMOTE, combined with augmented features, improves ensemble models the most. In addition, the ablation study confirms that K-Means SMOTE and the ET classifier have the most significant impact, with incremental gains offered by IRFs. The whole configuration is validated as the optimal setup for this study.

## 5.4 The Findings

The ET model, using the combined dataset, applied K-Means SMOTE and added 4 IRFs, was shown to be the best option based on the assessment criteria of accuracy, precision, recall, F1-score, and AUC-ROC. Maintaining a well-balanced trade-off between precision and recall, ET commonly delivered good performance across all measures, with the best accuracy (96.47%), substantial precision (94.79%), and solid recall (97.86%), reflected in a high F1-score (96.30%). Its AUC-ROC (99.51%) was also the greatest, suggesting exceptional classification capability. Although specific models, such as RF and GB, also showed good performance, ET's stability and robustness over all measures made it the best one. In comparison, the 4 IRFs with K-Means SMOTE enhance model performance more effectively, while the 8 IRFs setup shows minor declines, suggesting added complexity may introduce noise. For this work, ET with K-Means SMOTE and 4IRFs is thus the most efficient model-data combination, as it strikes the optimal balance between predictive power and generalization.

Among ensemble models, the ET model outperforms RF and GB due to its unique approach to randomness and feature selection. ET uses a different splitting technique than RF and GB. It chooses split points entirely at random, while RF determines the optimal split depending on impurity reduction, and GB creates trees consecutively to reduce errors. This additional randomization in ET helps prevent overfitting to synthetic data from K-Means SMOTE, which RF may find challenging due to its more deterministic character. Furthermore, IRF provides more information for classification, and ET is ideal for effectively utilizing high-dimensional data, as it does not rely on boosting like GB, which can be sensitive to noise in augmented features. While RF and GB show outstanding performance, they are more prone to overfitting or being sensitive to synthetic data and augmented features. In contrast, ET remains robust, achieving the maximum accuracy, recall, and AUC-ROC, thereby making it the best model for this dataset combination.

Due to their inherent model characteristics and susceptibility to synthetic data and feature expansion, SVM, ANN, and KNN do not significantly benefit from the combined dataset. SVM relies on determining an ideal hyperplane for classification; however, the inclusion of K-Means SMOTE synthetic samples and additional features can compromise the margin and result in a minor precision-recall trade-off. KNN suffers the most since it is susceptible to high-dimensional spaces, and adding IRF increases feature dimensionality, thereby weakening distance-based classification by introducing irrelevant or duplicate information. K-Means SMOTE also creates new samples through interpolation, which may distort KNN's closest neighbor computations and lead to performance variations. Although usually characterized by strong to complicated feature interactions, ANN does not demonstrate notable benefits, as it relies on considerable hyperparameter adjustment to effectively learn from synthetic and augmented data, which may not be sufficiently optimal in this case. These individual learners fail to identify meaningful patterns from synthetic and augmented data, thereby restricting their performance gains. In contrast, ensemble-based models, such as ET and RF, efficiently handle noise and utilize feature diversity. Especially, the ET model tends to be the best option based on overall performance.

By applying medical domain knowledge, developing new features, and ensuring that data inputs align with established clinical reasoning, IRF significantly enhances the explainability and trustworthiness of disease prediction models. As they relate to well-known physiological and pathological concepts, models incorporating medically meaningful features, such as risk scores, biomarker ratios, and symptom-based indices, enable healthcare professionals to interpret predictions more clearly. This transparency enables healthcare professionals to validate model decisions, thereby enhancing their trust in predictions generated by machine learning. Moreover, domain-specific feature engineering reduces the “black-box” characteristics of ML, thereby facilitating the identification of the reasons a patient is classified as high or low risk. In healthcare, trustworthy artificial intelligence requires that judgments be reasonable to both healthcare providers and patients, interpretable in line with established medical best practices, and aligned with these practices. Expert knowledge enhances model credibility, boosting clinical adoption and patient trust in AI-driven diagnosis and treatment.

## 5.5 Limitations

This study was conducted using retrospective EHR data from a single hospital, which may limit generalizability. The dataset lacked neuropsychological and imaging features, restricting the scope of prediction. Additionally, while IRFs enhanced interpretability, they may not capture all clinical nuances without expert input, as Real-world clinical validation and prospective studies were not conducted. Although promising results are obtained, this study has several limitations. The proposed approach yielded the best performance in ensemble-based models, such as Extra Trees, but showed limited improvements in SVM, KNN, and ANN, suggesting model-specific effectiveness. The use of K-Means SMOTE may also distort local data structure, particularly affecting distance-based or margin-based models. Increasing IRFs beyond a certain point introduced noise and degraded performance due to dimensionality issues. ANN models underperformed, likely due to insufficient hyperparameter tuning. Moreover, the dataset's origin from a single clinical setting raises concerns about generalizability, and the lack of validation from clinical users limits confirmation of the model's practical interpretability and usability.

## 5.6 Suggestion and Future Study

The proposed model is well-suited for disease prediction tasks involving imbalanced, feature-rich data, providing strong generalization without overfitting. Synthetic augmentation and ensemble methods boost accuracy and trust, supporting early diagnosis, risk assessment, and personalized medicine. Future work should focus on optimizing deep learning models through advanced hyperparameter tuning and targeted feature selection to maximize the utility of synthetic and augmented data. To further enhance model performance and generalizability, more sophisticated data generation techniques, such as generative adversarial networks or adaptive resampling, should be explored as alternatives to K-Means SMOTE. In addition, future studies should evaluate the model's real-time diagnostic support within clinical workflows and investigate its integration into mobile health applications and national health information systems to improve practical deployment and interoperability.

Moreover, although this study focused on binary classification to distinguish dementia cases from non-dementia, the proposed approach can be extended to multiclass classification involving various dementia subtypes. This is feasible when the dataset remains imbalanced and contains only interrelated clinical features. The inter-relation-based feature (IRFs) was designed in a disease-independent manner, making it adaptable not only to different forms of dementia but also to other medical conditions that share similar dataset characteristics. The model's ability to extract meaningful relationships between features and manage class imbalance enhances its applicability across a wide range of diagnostic tasks. Furthermore, the framework emphasizes model transparency and sensitivity to subtle risk factors—critical elements in clinical Diagnostic decision-making.

## 5.7 Conclusion

This study proposes the ET model and a data enrichment method for dementia classification, thereby enhancing the early detection of dementia. The proposed data enrichment method was a hybrid approach that combined feature augmentation and data balancing to enhance the data dimension, provide more informative features, and ensure a sufficient number of samples from the minority class. For feature augmentation, Inter-Relation-based Features (IRF) were proposed, leveraging medical domain knowledge to promote the explainability and trustworthiness of the model. The K-Means SMOTE was applied as a method to handle imbalanced data by generating new data based on the actual clusters of the original dataset. Consequently, the original dataset was transformed into a higher-dimensional space, making it suitable for model construction. The study utilized 14,763 EHR records and an initial set of 22 features from a hospital in Chiang Rai, Thailand. The ET model was proposed for classification due to its ability to assess feature importance and handle multicollinearity. The model's performance was compared to other traditional and ensemble learning methods. Experimental results demonstrated that the combination of 4 IRFs and K-Means SMOTE significantly enhanced the performance of the ET model across various metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

Future research should adapt its application to distinguish between dementia severity levels or in differential diagnosis of other conditions, as well as integrate it with more advanced synthetic sampling techniques for better generalization and dimensionality, and put it into practical application.

# REFERENCES

Aashima, Bhargav, S., Kaushik, S., & Dutt, V. (2021). A combination of decision trees with machine learning ensembles for blood glucose level predictions. In M. Saraswat, S. Roy, C. Chowdhury & A. H. Gandomi (Eds.), *Proceedings of International Conference on Data Science and Applications* (vol. 287, pp 533–548). Springer, Singapore. https://doi.org/10.1007/978-981-16-5348-3_42

Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications, 10*(8), 14–23. https://doi.org/10.14569/IJACSA.2019.0100813

Almubark, I., Alsegehy, S., Jiang, X., & Chang, L. C. (2020). Early detection of mild cognitive impairment using neuropsychological data and machine learning techniques. In *2020 IEEE Conference on Big Data and Analytics (ICBDA)* (pp. 32–37). IEEE. https://doi.org/10.1109/ICBDA50157.2020.9289741

Bansal, D., Khanna, K., Chhikara, R., Dua, R. K., & Malhotra, R. (2022). Comparative analysis of artificial neural networks and deep neural networks for detection of dementia. *International Journal of Social Ecology and Sustainable Development (IJSESD), 13*(9), 1–18. https://doi.org/10.4018/IJSESD.313966

Beebe-Wang, N., Okeson, A., Althoff, T., & Lee, S. I. (2021). Efficient and explainable risk assessments for imminent dementia in an aging cohort study. *Journal of Biomedical and Health Informatics, 25*(7), 2409–2420. https://doi.org/10.1109/JBHI.2021.3059563

Bishop, M. L., Fody, E. P., & Schoeff, L. E. (2023). *Clinical chemistry: Principles, techniques, and correlations* (8th ed.). Wolters Kluwer.

Castellazzi, G., Cuzzoni, M. G., Ramusino, M. C., Martinelli, D., Denaro, F., Ricciardi, A., . . . Wheeler-Kingshott, C. A. M. G. (2020). A machine learning approach for the differential diagnosis of Alzheimer and vascular dementia fed by MRI selected features. *Frontiers in Neuroinformatics, 14*, 25. https://doi.org/10.3389/fninf.2020.00025

Chaiyo, Y., Rueangsirarak, W., Hristov, G., & Temdee, P. (2025). Improving early detection of dementia: Extra trees-based classification model using inter-relation-based features and k-means synthetic minority oversampling technique. *Big Data Cogn. Comput, 9*(6), 148. https://doi.org/10.3390/bdcc90 60148

Cura, O. K., Yilmaz, G. C., Ture, H. S., & Akan, A. (2022). Deep time-frequency feature extraction for Alzheimer's dementia EEG classification. In *2022 Medical Technologies Congress (TIPTEKNO)* (pp. 1-4). IEEE. https://doi.org/10.1109/TIPTEKNO56568.2022.9960155

Fern´andez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *Journal of Machine Learning Research, 15*(1), 3133–3181. https://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf

Fernández, A., García, S., Galar, M., Prati, R. C. Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data streams. In *Learning from imbalanced data sets* (pp. 279-303). *Springer*. https://doi.org/10.1007/978-3-319-98074-4

Goel, A., Lal, M., & Javadekar, A. N.  (2023). Comparative analysis of the machine and deep learning classifier for dementia prediction. In 2023 *Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)* (pp. 1-8). IEEE. https://doi.org/10.1109/ACCTHPA57160.2023.10083361

Gómez, C., Vaquerizo-Villar, F., Poza, J., Ruiz, S. J., Tola-Arribas, M. A., & Cano, M. (2017). Bispectral analysis of spontaneous EEG activity from patients with moderate dementia due to Alzheimer's disease. In 2017 *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 422-425). IEEE. https://doi.org/10.1109/EMBC.2017.8036852

Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., . . . van der Flier, W. M. (2023). Global estimates on the number of persons across the Alzheimer's disease continuum. *Alzheimers & Dementia, 19*(2), 658-670. https://doi.org/10.1002/alz.12694

Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes. *Journal Teknologi dan Sistem Komputer, 8*, 89-93. https://doi: 10.14710/jtsiskom.8.2.2020.89-93

Hall, J. E. (2021). *Guyton and Hall textbook of medical physiology* (13th ed.). Saunders.

Hanai, S., Kato, S., Sakuma, T., Ohdake, R., Masuda, M., & Watanabe, H. (2022). A dementia classification based on speech analysis of casual talk during a clinical interview. In 2022 *IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)* (pp. 38- 40). https://doi.org/10.1109/LifeTech53646.2022.9754933

Hanczár, G., Stippinger, M., Hanák, D., Kurbucz, M. T., Törteli, O. M., Chripkó, Á., . . . Somogyvári, Z. (2023). Feature space reduction method for ultrahigh-dimensional, multiclass data: Random forest-based multiround screening (RFMS). *Machine Learning: Science and Technology, 4*(4). https://doi.org/10.1088/2632-2153/ad020e

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE. https://doi.org/10.1109/IJCNN.2008.4633969

Jameson, J. L., Fauci, A. S., Kasper, D. L., Hauser, S. L., Longo, D. L., & Loscalzo, J. (2022). *Harrison's principles of internal medicine* (21st ed.). McGraw Hill. https://accessmedicine.mhmedical.com/book.aspx?bookid=3541

Javeed, A., Dallora, A. L., Berglund, J. S., Idrisoglu, A., Ali, L., Rauf, H. T., . . . Anderberg, P. (2023). Early prediction of dementia using feature extraction battery (FEB) and optimized support vector machine (SVM) for classification. *Biomedicines, 11*(2), 439. https://doi.org/10.3390/biomedicines11020439

Jeong, J., Chae, J. H., Kim, S. Y., & Han, S. H. (2001). Nonlinear dynamic analysis of the EEG in patients with Alzheimer's disease and vascular dementia. *Journal of Clinical Neurophysiology, 18*(1), 58-67. https://doi.org/10.1097/00004691-200101000-00010

Jha, A., John, E., & Banerjee, T. (2022). Multi-class classification of dementia from MRI images using transfer learning. In *2022 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 0597-0602). IEEE. http://doi.org/10.1109/UEMCON54665.2022.9965672

Justin, B. N., Turek, M., & Hakim, A. M. (2013). Heart disease as a risk factor for dementia. *Clinical Epidemiology, 5*, 135-145. https://doi.org/10.2147/CLEP.S30621

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kramer, O. (2013). *Dimensionality reduction with unsupervised nearest neighbors*. Springer. https://doi.org/10.1007/978-3-642-38652-7

Kunapuli, G. (2023). *Ensemble methods for machine learning*. Simon and Schuster.

Lastuka, A., Bliss, E., Breshock, M. R., Iannucci, V. C., Sogge, W., Taylor, K. V., . . . Dieleman, J. L. (2024). Societal costs of dementia: 204 countries, 2000–2019. *Journal of Alzheimer's Disease, 101*(1), 277-292. https://doi.org/10.3233/JAD-240163

Liu, Q. S., Xue, Y., Li, G., Qiu, D., Zhang, W., Guo, Z., . . . Li, Z. (2023). Application of KM-SMOTE for rockburst intelligent prediction. *Tunnelling and Underground Space Technology, 138*, 105180. https://doi.org/10.1016/j.tust.2023.105180

Mirzaei, G., & Adeli, H. (2022). Machine learning techniques for diagnosis of Alzheimer's disease, mild cognitive disorder, and other types of dementia. *Biomedical Signal Processing and Control, 72*, 103293. https://doi.org/10.1016/j.bspc.2021.103293

Mohammed, B. A., Senan, E. M., Rassem, T. H., Makbol, N. M., Alanazi, A. A., Al-Mekhlafi, Z. G., . . . Ghaleb, F. A. (2021). Multi-method analysis of medical records and MRI images for early diagnosis of dementia and Alzheimer's disease based on deep learning and hybrid methods. *Electronics, 10*(22), 2860. https://doi.org/10.3390/electronics10222860

Muangpaisan, W. (2013). *Dementia: Prevention, assessment and care*. Parbpim.

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array, 16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Nancy, A., Balamurugan, M., & Vijaykumar, S. (2017). A brain EEG classification system for the mild cognitive impairment analysis. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICACCS.2017.8014655

Narasimhan, R., Gopalan, M., Sikkandar, M. Y., Alassaf, A., AlMohimeed, I., Alhussaini, K., . . . Sheik, S. B. (2021). Employing deep-learning approach for the early detection of mild cognitive impairment transitions through the analysis of digital biomarkers. *Sensors, 23*(21). https://doi.org/10.3390/s23218867

Narmatha, C., Alatawi, H., & Alatawi, H. Q. (2021). An analysis of deep learning techniques in neuroimaging. *Journal of Computational Science and Intelligent Technologies, 2*(1), 7-13.

Nichols, E., Steinmetz, J. D., Vollset, F S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., . . . Vos, T. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health, 7*(2), e105-e125.

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology, 24(*12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Öcal, H. (2024). A novel approach to detection of Alzheimer’s disease from handwriting: Triple ensemble learning model. *Gazi University Journal of Science Part C: Design and Technology, 12*(1), 214-223. https://doi.org/10.29109/gujsc.1386416

Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019). K-nearest neighbor (k-NN) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 83-88). IEEE. https://doi.org/10.1109/ICSITech46713.2019.8987530

Pritchard, W. S., Duke, D. W., Coburn, K. L., Moore, N. C., Tucker, K. A., Jann, M. W., . . . Hostetler, R. M. (1994). EEG-based, neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by non-linear EEG measures. *Electroencephalography and Clinical Neurophysiology, 91*(2), 118-130. https://doi.org/10.1016/0013-4694(94)90033-7

Ranjan, N., Kumar, D. U., Dongare, V., Chavan, K., & Kuwar, Y. (2022). Diagnosis of Parkinson disease using handwriting analysis. *International Journal of Computer Applications, 184*(1), 13-16.

Reddy, T. S., Saikiran, V., Samhitha, S., Moin, S., Kumar, T. P., & Charan, V. S. (2023). Early detection of Alzheimer's disease using data augmentation and CNN. In 2023 *4th IEEE Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE. https://doi.org/10.1109/GCAT59970.2023.10353397

Rodrigues, P. M., Bispo, B. C., Freitas, D. R., Teixeira, J. P., & Carreres, A. (2013). Evaluation of EEG spectral features in Alzheimer disease discrimination. In *21st European Signal Processing Conference (EUSIPCO 2013)* (pp. 1-5). IEEE. https://ieeexplore.ieee.org/document/6811669

Salinas Ruíz, J., Montesinos López, O. A., Hernández Ramírez, G., & Crossa Hiriart, J. (2023). Generalized linear models. In *Generalized linear mixed models with applications in agriculture and biology* (pp. 43-84). Springer. https://doi.org/10.1007/978-3-031-32800-8_2

Samanta, S., Mazumder, I., & Roy, C. (2023). Deep learning-based early detection of Alzheimer's disease using image enhancement filters. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICAECT57570.2023.10117880

Shafique, R., Mehmood, A., & Choi, G. S. (2019). *Cardiovascular disease prediction system using extra trees classifier* [Preprint]. *Research Square*, 11, 51. https://doi.org/10.21203/rs.2.14454/v1

Shen, Y., Zhu, J., Deng, Z., Lu, W., & Wang, H. (2023). EnsDeepDP: An ensemble deep learning approach for disease prediction through metagenomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20*(2), 986-998. https://doi.org/10.1109/TCBB.2022.3201295

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*, 60. https://doi.org/10.1186/s40537-019-0197-0

Skerrett, P. J. (2014). *Lipid disorders: Diagnosis and treatment*. Harvard Health.

Thongwachira, C., Jaignam, N., & Thophon, S. (2019). A model of dementia preventi on in older adults at Taling Chan District Bangkok Metropolis. *KKU Research Journal (Graduate Studies), 19*(3), 96-108. https://ph02.tci-thaijo.org/index.php/gskku/article/view/208436

Trambaiolli, L. R., Spolaôr, N., Lorena, A. C., Anghinah, R., & Sato, J. R. (2017). Feature selection before EEG classification supports the diagnosis of Alzheimer’s disease. *Clinical Neurophysiology, 128*(10), 2058-2067. https://doi.org/10.1016/j.clinph.2017.06.251

Ullah, H. M. T., Onik, Z., Islam, R., & Nandi, D. (2018). Alzheimer’s disease and dementia detection from 3D brain MRI data using deep convolutional neural networks. In *2018 3rd International Conference for Convergence in Technology (I2CT)* (pp. 1-3). https://doi.org/10.1109/I2CT.2018.8529808

Vardhini, K. V., Vishnumolakala, L. D., Palanki, S. U. A., Yarramsetty, M., & Raja, G. (2024). Alzheimer’s research and early diagnosis through improved deep learning models. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 1577-1583). IEEE. https://doi.org/10.1109/ICESC60852.2024.10689869

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., . . . Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation. *Medical Image Analysis, 63*, 101694. https://doi.org/10.1016/j.media.2020.101694

World Health Organization. (2019). *Risk reduction of cognitive decline and dementia: WHO guidelines*. https://www.who.int/publications/i/item/risk-reduction-of-cognitive-decline-and-dementia

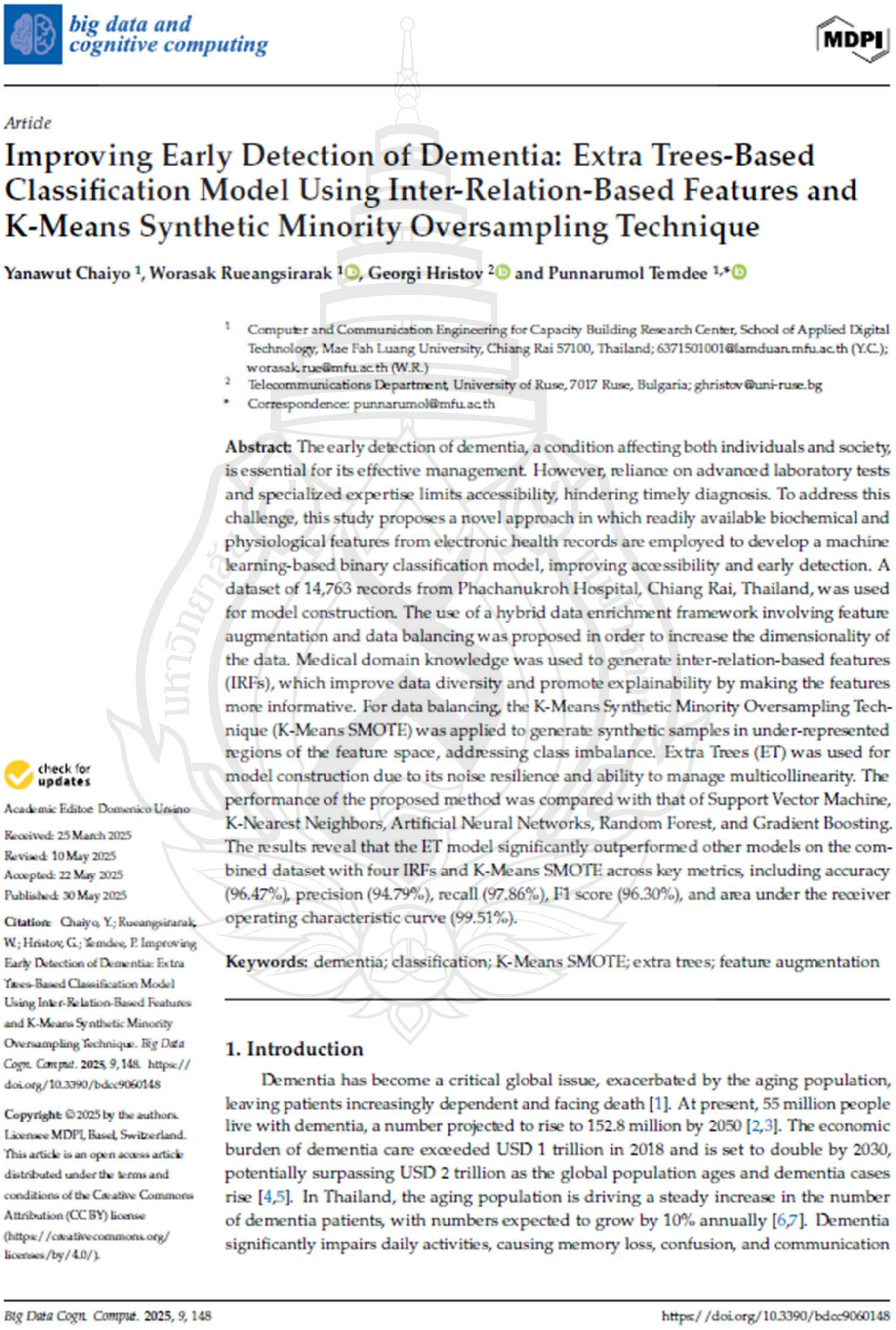
Yongcharoenchaiyasit, K., Arwatchananukul, S., Temdee, P., & Prasad, R. (2023). Gradient boosting-based model for elderly heart failure, aortic stenosis, and dementia classification. *IEEE Access, 11*, 48677-48696. https://doi.org/10.1109/ACCESS.2023.3276468

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval, 14*(1), 1-101. http://dx.doi.org/10.1561/1500000066

Zhou, Z.-H., & Feng, J. (2017). Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 3553-3559). IJCAI. https://doi.org/10.24963/ijcai.2017/497

# APPENDIX A

# PUBLICATIONS



big data and cognitive computing

MDPI

*Article*

## Improving Early Detection of Dementia: Extra Trees-Based Classification Model Using Inter-Relation-Based Features and K-Means Synthetic Minority Oversampling Technique

**Yanawut Chaiyo [1], Worasak Rueangsirarak [1], Georgi Hristov [2] and Punnarumol Temdee [1,*]**

[1] Computer and Communication Engineering for Capacity Building Research Center, School of Applied Digital Technology, Mae Fah Luang University, Chiang Rai 57100, Thailand; 6371501001@lamduan.mfu.ac.th (Y.C.); worasak.rue@mfu.ac.th (W.R.)

[2] Telecommunications Department, University of Ruse, 7017 Ruse, Bulgaria; ghristov@uni-ruse.bg

[*] Correspondence: punnarumol@mfu.ac.th

**Abstract:** The early detection of dementia, a condition affecting both individuals and society, is essential for its effective management. However, reliance on advanced laboratory tests and specialized expertise limits accessibility, hindering timely diagnosis. To address this challenge, this study proposes a novel approach in which readily available biochemical and physiological features from electronic health records are employed to develop a machine learning-based binary classification model, improving accessibility and early detection. A dataset of 14,763 records from Phachanukroh Hospital, Chiang Rai, Thailand, was used for model construction. The use of a hybrid data enrichment framework involving feature augmentation and data balancing was proposed in order to increase the dimensionality of the data. Medical domain knowledge was used to generate inter-relation-based features (IRFs), which improve data diversity and promote explainability by making the features more informative. For data balancing, the K-Means Synthetic Minority Oversampling Technique (K-Means SMOTE) was applied to generate synthetic samples in under-represented regions of the feature space, addressing class imbalance. Extra Trees (ET) was used for model construction due to its noise resilience and ability to manage multicollinearity. The performance of the proposed method was compared with that of Support Vector Machine, K-Nearest Neighbors, Artificial Neural Networks, Random Forest, and Gradient Boosting. The results reveal that the ET model significantly outperformed other models on the combined dataset with four IRFs and K-Means SMOTE across key metrics, including accuracy (96.47%), precision (94.79%), recall (97.86%), F1 score (96.30%), and area under the receiver operating characteristic curve (99.51%).

**Keywords:** dementia; classification; K-Means SMOTE; extra trees; feature augmentation

check for updates

Academic Editoe: Domenico Ursino

Received: 25 March 2025
Revised: 10 May 2025
Accepted: 22 May 2025
Published: 30 May 2025

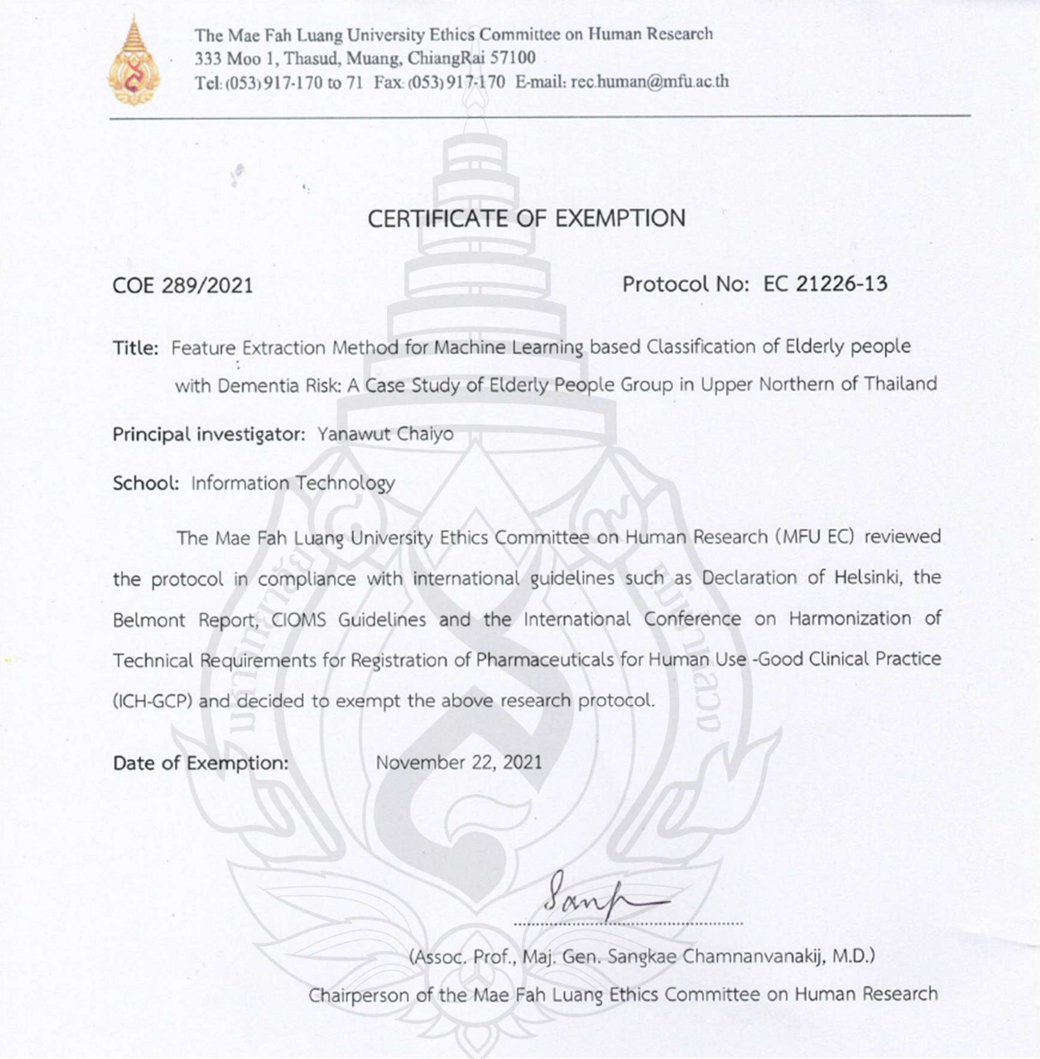
**Citation:** Chaiyo, Y.; Rueangsirarak, W.; Hristov, G.; Temdee, P. Improving Early Detection of Dementia: Extra Trees-Based Classification Model Using Inter-Relation-Based Features and K-Means Synthetic Minority Oversampling Technique. *Big Data Cogn. Comput.* **2025**, *9*, 148. https://doi.org/10.3390/bdcc9060148

### 1. Introduction

Dementia has become a critical global issue, exacerbated by the aging population, leaving patients increasingly dependent and facing death [1]. At present, 55 million people live with dementia, a number projected to rise to 152.8 million by 2050 [2,3]. The economic burden of dementia care exceeded USD 1 trillion in 2018 and is set to double by 2030, potentially surpassing USD 2 trillion as the global population ages and dementia cases rise [4,5]. In Thailand, the aging population is driving a steady increase in the number of dementia patients, with numbers expected to grow by 10% annually [6,7]. Dementia significantly impairs daily activities, causing memory loss, confusion, and communication

# APPENDIX B

# ETHICAL APPROVAL CERTIFICATE



The Mae Fah Luang University Ethics Committee on Human Research
333 Moo 1, Thasud, Muang, ChiangRai 57100
Tel: (053) 917-170 to 71 Fax: (053) 917-170 E-mail: rec.human@mfu.ac.th

## CERTIFICATE OF EXEMPTION

COE 289/2021 Protocol No: EC 21226-13

**Title:** Feature Extraction Method for Machine Learning based Classification of Elderly people with Dementia Risk: A Case Study of Elderly People Group in Upper Northern of Thailand

**Principal investigator:** Yanawut Chaiyo

**School:** Information Technology

The Mae Fah Luang University Ethics Committee on Human Research (MFU EC) reviewed the protocol in compliance with international guidelines such as Declaration of Helsinki, the Belmont Report, CIOMS Guidelines and the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use -Good Clinical Practice (ICH-GCP) and decided to exempt the above research protocol.

**Date of Exemption:** November 22, 2021

.............................................................

(Assoc. Prof., Maj. Gen. Sangkae Chamnanvanakij, M.D.)
Chairperson of the Mae Fah Luang Ethics Committee on Human Research

# CURRICULUM VITAE

| | |
|---|---|
| **NAME** | Yanawut Chaiyo |
| **EDUCATIONAL BACKGROUND** | |
| 2008 | Bachelor of Industrial Technology Computer<br>Industrial Technology<br>Chiang Rai Rajabhat University |
| 2016 | Master of Computer Engineering<br>Computer Engineering<br>Chiang Mai University |
| **WORK EXPERIENCE** | |
| 2019-Present | System Analysis (SA)<br>Summit Computer |