EXTRA TREES MODEL WITH MINORITY TARGET OVERSAMPLING FOR CLASSIFICATION OF DEMENTIA AND HEART FAILURE IN ADULTS

PORNTHEP PHANBUA

MASTER OF ENGINEERING
IN
COMPUTER ENGINEERING

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY
2024

EXTRA TREES MODEL WITH MINORITY TARGET OVERSAMPLING FOR CLASSIFICATION OF DEMENTIA AND HEART FAILURE IN ADULTS

PORNTHEP PHANBUA

THIS THESIS IS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING IN COMPUTER ENGINEERING

SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY

2024

THESIS APPROVAL

MAE FAH LUANG UNIVERSITY

FOR

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

Thesis Title: Extra Trees Model with Minority Target Oversampling for Classification of Dementia and Heart Failure in Adults

Author: Pornthep Phanbua

Examination Committee:

Associate Professor Adisorn Leelasantitham, Ph. D.	Chairperson
Associate Professor Punnarumol Temdee, Ph. D.	Member
Assistant Professor Sujitra Arwatchananukul, Ph. D.	Member
Associate Professor Nattapol Aunsri, Ph. D.	Member
Surapong Uttama, Ph. D.	Member

Advisors:

P. Temdee .....Advisor

(Associate Professor Punnarumol Temdee, Ph. D.)

Sujitra A. .....Co-Advisor

(Assistant Professor Sujitra Arwatchananukul, Ph. D.)

Dean:

.....

(Assistant Professor Nacha Chondamrongkul, Ph. D.)

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who supported me throughout this thesis.

First and foremost, I sincerely thank my advisor, Assoc. Prof. Punnarumol Temdee, Ph.D., and my co-advisor, Asst. Prof. Sujitra Arwatchananukul, Ph.D., for their invaluable guidance, support, and the opportunity to work under their mentorship. Their patience, expertise, and encouragement have been crucial in shaping this work.

I also want to thank my colleagues for their camaraderie and the stimulating discussions we shared. Their support and collaboration have been a great source of inspiration.

Finally, I am deeply grateful to my family for their patience, understanding, and constant encouragement, which kept me motivated throughout this journey.

Pornthep Phanbua

Thesis Title	Extra Trees Model with Minority Target Oversampling for Classification of Dementia and Heart Failure in Adults
Author	Pornthep Phanbua
Degree	Master of Engineering (Computer Engineering)
Advisor	Assoc. Prof. Punnarumol Temdee, Ph. D.
Co-Advisor	Asst. Prof. Sujitra Arwatchananukul, Ph. D.

ABSTRACT

Recent advancements in medical technology have led to increased longevity, resulting in aging populations across many nations. Research indicates that experiencing coronary heart disease at a young age can have a significantly negative impact on brain health. Specifically, individuals who suffer from heart failure are found to be 60% more likely to develop dementia. This study proposes a new early detection method for dementia by distinguishing it from the associated disease of heart failure using blood testing data. In particular, this study uses extra trees (ETs) to create a classification model for identifying individuals with dementia and heart failure based on personal and clinical data collected during their routine health checkups. The study utilizes a dataset comprising 4,297 records from Chiang Rai Prachanukroh Hospital in Chiang Rai Province, Thailand. The performance of the proposed model is evaluated and compared with other methods, including decision tree, K-nearest neighbors, support vector machine, random forest, gradient boosting, and adaptive boosting. The results demonstrate that the proposed ET model outperforms other methods in terms of accuracy (89.11%), precision (94.92%), recall (92.50%), F-measure (93.70%), and ROC AUC score (78.88%).

Keywords: Dementia, Heart Failure, Prediction, Machine Learning, Extra Trees, Classification

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Problems

The world is currently facing the challenges of an aging population. In 2017, the number of individuals aged 60 or older reached 962 million, which is more than double the figure of 392 million in 1980. By 2050, this number is projected to double again, reaching nearly 2.1 billion. This would surpass the population of adolescents aged 10-24, which is 2.0 billion [1]. Aging is a natural process that is associated with a higher likelihood of developing various diseases. The changes that come with age can be categorized into normal aging, common diseases, functional changes, cognitive/psychiatric changes, and social changes [2]. This study primarily focuses on examining cognitive changes and common diseases. The domain of cognitive changes in aging includes alterations in cognitive abilities such as memory, attention, and reasoning. Dementia is a condition that falls within this domain. A study has shown a clear and consistent correlation between the age at which heart disease begins and the likelihood of developing dementia [3]. It is important to pay particular attention to individuals diagnosed with heart disease at a young age, as they may be at a higher risk of dementia. According to the literature, heart failure has also been identified as a significant risk factor for dementia, with a 60% increased risk observed in individuals with heart failure compared to those without [4]. Furthermore, the study revealed that the overall prevalence of cognitive impairment and dementia in all heart disease patients was 41.42% and 19.79%, respectively [5]. Given the well-documented correlation between heart failure and dementia, it is crucial to develop a classification model that can effectively differentiate between these two conditions. This model will play a significant role in enabling early detection and treatment, thereby improving patient outcomes and reducing healthcare expenses.

Dementia is a disorder that directly affects the brain, impairing the patient's cognitive abilities, particularly memory. It also significantly impacts language, attention, decision-making, and planning [6]. In 2015, an estimated 47 million people worldwide were living with dementia, and this number is projected to triple by 2050 [7]. Diagnosis involves assessing medical history, conducting cognitive and physical examinations, laboratory tests, and brain imaging. Managing dementia requires a combination of non-pharmacological and pharmacological methods, although the effectiveness of existing treatments is still limited [8]. Currently, there is no definitive treatment for dementia [9]. However, certain medications inhibit an enzyme responsible for breaking down acetylcholine, a neurotransmitter crucial for cholinergic neurons in both the peripheral and central nervous systems, highlighting its significance. Therefore, therapies that aim to replace lost neurons and prevent neuronal death have the potential to modify the progression of dementia and alleviate its symptoms [10]. Furthermore, it is crucial for patients to manage risk factors that increase the likelihood of a stroke, such as high blood pressure and diabetes [6].

According to a report by the Heart Failure Society of America, approximately 6.5 million individuals aged 20 and above in the United States suffer from heart failure, with around 960,000 new cases occurring each year. Additionally, 8.5% of deaths related to heart disease can be directly linked to heart failure [11]. Heart failure is considered the final stage of various heart conditions, including cardiomyopathy, valvular or ischemic heart disease, and others [12]. If not treated, these conditions pose a significant risk of repeated hospitalization and death [13]. Heart failure is a prevalent condition in elderly patients, and its occurrence is influenced by multiple factors associated with aging, such as structural, biochemical, clinical, and psychological factors [14].

Common symptoms of heart failure include fatigue, dyspnea, swollen ankles, exercise intolerance, or symptoms related to the underlying cause. However, it may not be enough to solely rely on the presence of clinical features for diagnosis, especially in women and elderly or obese patients [15]. Heart failure is a chronic condition that cannot be cured. Currently, there are four main treatment options available, which are tailored to the patient's symptoms: making healthy lifestyle changes, using implantable devices for heart rhythm control, taking medications, and undergoing surgical interventions.

These treatment modalities have significantly improved both the survival and quality of life of patients [16].

With the rapid advancement of technology, the amount of detail included in a patient's medical information has significantly increased. This includes clinical data, diagnosis, and laboratory examination results. However, this data often varies in structure and format across different hospitals, and errors may occur due to inaccurate or incomplete data collection. Despite these challenges, machine learning methods have successfully utilized this medical data, resulting in the development of numerous valuable applications in the medical and healthcare fields. Many previous studies have focused on using medical data to classify whether a patient has a specific disease, without considering the risk of concurrent diseases [17–19]. In contrast, this study aims to develop a predictive model for dementia by employing a binary classification between dementia and heart failure, two highly correlated diseases. The proposed model utilizes ensemble learning, specifically extra trees (ETs), to improve the performance of individual models by combining their predictions. The ultimate goal of this classification model is to provide patients with awareness of their risk for both dementia and heart failure. This knowledge can then facilitate timely treatment interventions.

1.2 Research Objectives

The objective of this study is to make a significant contribution to the early prediction of high-risk diseases that are commonly observed among the elderly population which are dementia and heart failure. By promptly identifying patients in the early stages of these diseases, individuals can proactively take measures to effectively manage their health. This approach promotes timely self-care and facilitates prompt access to treatment, thus alleviating the financial burden associated with expensive healthcare costs. Additionally, the model developed through this study has the potential to support specialists in conducting preliminary risk assessments and can provide valuable assistance in making informed decisions.

1.3 Scope

This study primarily focuses on developing machine learning models specifically designed to differentiate between dementia and heart failure by employing effective algorithms and techniques to enhance the accuracy and reliability of the models. Then, evaluate the performance of the developed models through rigorous testing and validation procedures.

1.4 Thesis Structure

The thesis structure is introduced in this section to outline the organization and sequence of the work. Each chapter is briefly summarized, highlighting the main content and objectives. This summary allows readers to gain a comprehensive overview of the thesis and grasp the key components and progression of the research.

Chapter 2: Literature Review, this chapter presents a comprehensive review of relevant literature, including previous studies and research findings, to establish the theoretical foundation and contextualize the research.

Chapter 3: Research Methodology, this chapter presents the detail of the research methodology including experimental procedures, data collection, data preprocessing, data modeling, and model evaluation to achieve the experiment result.

Chapter 4: Results, this chapter presents the results of the research study, which included the Pearson correlation coefficient, accuracy, precision, recall, F1-score, and AUC-ROC score. These measures were used to assess the performance and predictive capabilities of the implemented model.

Chapter 5: Conclusion, this chapter synthesizes the outcomes of this study, offering information into their implications and providing recommendations for future research directions.

CHAPTER 2

LITERATURE REVIEWS

This section provides an overview of current research on the development of a classification model for dementia and heart failure. With the continuous advancement of technology, the amount of available data is increasing, and machine learning classifiers are increasingly being used to analyze and predict the likelihood of these diseases. This section aims to provide a thorough understanding of the current state of research in this field.

2.1 Related Work

2.1.1 Dementia and Heart Failure Association

Heart failure increases the risk of developing dementia due to reduced blood flow caused by the heart's inefficiency. This diminished circulation can lead to cerebral ischemia, a condition that impacts the brain's blood vessels, known as cerebral blood flow, ultimately leading to their deterioration. As a result, this process triggers cognitive problems related to the accumulation of tau protein and decreased levels of amyloid-β in the plasma [20], both of which have been linked to predicting the risk of dementia [21]. The decline in cerebral blood flow appears to worsen cognitive impairment associated with heart failure, while individuals who undergo heart transplantation often experience improved cerebral blood flow, resulting in enhanced cognitive function [22–23].

Numerous studies have highlighted the complex relationship between these conditions, revealing that individuals dealing with heart failure have a significantly higher chance of experiencing cognitive decline and dementia compared to those without heart-related problems. For example, a comprehensive study [24] found a higher occurrence of dementia among individuals with heart failure compared to those without.

Similarly, Wolters et al. [4] demonstrated the significant impact of heart failure as a major risk factor for dementia. The research revealed a staggering 60% increased risk of developing dementia in individuals diagnosed with heart failure compared to those without this heart condition, aligning with the higher occurrence of dementia among individuals with heart failure observed in an earlier study. This statistical finding emphasizes the substantial influence of heart failure on the heightened vulnerability to cognitive decline and potential development of dementia in affected individuals. Such evidence suggests the need for early detection of both dementia and heart failure, which is the objective of this study.

2.1.2 Predictive Model for Dementia

Researchers are continuously evolving and exploring predictive models for dementia, with many studies using data from direct magnetic resonance imaging (MRI) scans [25–31]. MRI is a crucial complementary imaging technique used to diagnose intracranial neurological injuries and predict cognitive and behavioral outcomes in patients [32]. In a study by [9], a model incorporating MRI data was developed, supplemented by the inclusion of clinical data from dementia patients. This integrated approach allowed for a more comprehensive model construction, leveraging both MRI data and clinically relevant information, potentially enhancing the accuracy and depth of insights into dementia-related research. However, the complexity and high cost associated with integrating MRI data into research present significant barriers to patient access. Moreover, recent literature has highlighted the extensive use of different machine-learning methods in developing predictive models for dementia. These methods encompass classical supervised learning techniques such as support vector machine (SVM) [25–26] and naïve Bayes [29], ensemble-based learning methods like gradient boosting (GB) [33] and random forest (RF) [34], as well as deep learning methods [29] designed to handle large datasets. Additionally, eXtreme gradient boosting (XGBoost) [9] has been recognized as an effective way to enhance model performance [27].

2.1.3 Predictive Model for Heart Failure

A review of the current studies on heart failure classification shows that there is variation in the data used, which can be attributed to differences in the sources of data acquisition. In a study by researchers in [35], they used a dataset that included six main

features: weight, systolic blood pressure, diastolic blood pressure, heart rate, oxygen saturation, and diuretic usage. Additionally, patient well-being information was collected through responses to eight questions. The XGBoost classifier was then used to predict cardiac decompensation events in patients with heart disease. In another study [36], RF was found to be the most accurate predictive model after evaluating several algorithms, using a dataset of 299 heart failure patients. Another study [37] divided the data into two groups for model construction, based on risk factors of heart disease. The first group consisted of patients with heart disease, influenced by unavoidable factors such as patient age and sex. The factors for the second group included lifestyle behaviors such as chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak, peak exercise slope, number of major vessels colored by fluoroscopy, and thallium scan. The SVM model was then applied to tune the hyperparameters to appropriately differentiate each dataset and merged to create a hybrid grid search algorithm (HGSA) with the highest accuracy.

2.1.4 Classification Model for Dementia and Heart Failure

In general, it is challenging to detect dementia in adults using MRI diagnosis [38]. Similarly, accurate examinations and treatment for heart failure require data collection devices [39]. Therefore, this study proposes the use of regular checkup data to develop classification models for these two diseases. Specifically, the blood test dataset and personal data are employed due to the significant collinearity among their features. Some of these features can help predict the occurrence of the diseases.

In the blood test dataset, hemoglobin plays a crucial role in maintaining cellular bioenergetic homeostasis by facilitating oxygen binding and transportation to tissues [40]. On the other hand, platelets are commonly used in the treatment of vascular injuries [41]. Both hemoglobin and platelet levels have been shown to correlate with disease activity [42–43], indicating their effectiveness in predicting disease progression [44]. Neutrophils and lymphocytes are types of white blood cells [45], and their ratio, known as the neutrophil-lymphocyte ratio, can serve as a biomarker providing insights into inflammatory activity during the acute phase of coronary syndrome [46]. Potassium, an electrolyte, has been suggested as a preventive measure against the onset of Alzheimer's disease when maintained at high levels of intake as a non-pharmacological therapy [47].

Creatinine and blood urea nitrogen, in conjunction with urine-specific gravity and osmolality, are used to assess kidney function and offer relatively insensitive indicators of kidney damage [48]. In terms of personal data, age and biological characteristics are closely linked to health-related behaviors [49] and are connected to the morbidity load of multimorbid patients [50]. When compared to traditional methodologies like MRI scans, which can be difficult to obtain and carry associated risks [51], particularly in resource-limited settings with limited access to advanced medical technology, using blood test datasets as diagnostic data offers a cost-effective and convenient alternative to conventional testing techniques [52]. The primary objective of this study is to investigate this alternative approach.

Many machine learning models have been extensively used for constructing models. Several previous studies have concentrated on popular binary classification models like K-nearest neighbors (KNN), decision tree (DT), and SVM. Furthermore, numerous studies have utilized techniques from ensemble learning groups, such as RF, GB, and adaptive boosting (AdaBoost). KNN, DT, and SVM have proven to be effective binary classification algorithms for decades. KNN classifies instances based on the majority class of their k-nearest feature space neighbors. It is particularly suitable for datasets with overlapping classes as it evaluates neighboring instances instead of identifying class domains [53]. DT, known for its intuitiveness, simplicity, accuracy, and prediction ability, is popular for prediction and categorization tasks [54]. Its effectiveness extends to targeting prevention and intervention for high-risk individuals [55]. On the other hand, SVMs are adaptable and effective classifiers even with small sample sets [56]. They have been successfully used in predicting diagnoses and prognoses for diseases such as Alzheimer's, schizophrenia, and depression [57].

RF is a sophisticated ensemble learning technique that randomly samples training datasets and aggregates predictions to construct DTs. This method reduces overfitting and effectively handles high-dimensional data, improving pattern identification [58–59]. GB utilizes multiple weak DTs instead of a single strong model to generate accurate predictions. By iteratively combining DTs that reflect residuals, it minimizes errors and enhances prediction accuracy, particularly as the number of iterations increases and the residuals decrease [60]. AdaBoost builds strong classifiers from weak ones through weighted voting. It achieves this by finding the optimal threshold in one dimension to

divide data into two categories. However, the algorithm may require improvements in certain scenarios due to its potential for poor generalization and overfitting [61]. While previous studies mainly focused on predicting a single disease using ensemble learning models, this analysis chooses the ETs algorithm because of its suitability for handling datasets with high collinearity among features, such as clinical data [62]. Additionally, the algorithm demonstrates robustness to outliers and effectively handles missing data, making it a fitting choice for this analysis. A comparative study [63] also showed that ET outperformed other ensemble learning classifiers in disease prediction tasks.

However, in terms of dataset, doctors often do not collect blood results from patients with dementia because there is no perceived necessity for such testing. This lack of data leads to an imbalanced dataset, a common issue in medical research where one class is underrepresented compared to others. Addressing imbalanced datasets is crucial for improving model performance, and several techniques have been developed for this purpose. One such method is Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC), which has been applied to various datasets, showing improved model performance by effectively balancing the classes, and has also been tested in real datasets like those related to heart failure, demonstrating its practical applicability and effectiveness [64-65]. Additionally, advanced techniques such as Conditional Tabular Generative Adversarial Networks (CTGAN) have been employed to generate high-quality synthetic data, effectively addressing the issue of imbalanced datasets. These synthetic data generation methods provide a robust solution for ensuring that machine learning models have sufficient representative data from all classes, thereby enhancing their predictive performance and reliability [66-68].

Therefore, this study proposes a new classifier model that uses the relationships between features in blood test data and personal data to classify two diseases: dementia and heart failure. Taking into account the correlation between the features in the blood test data and personal data, the ET classifier was chosen to model the datasets collected from Chiang Rai Prachanukroh Hospital in Chiang Rai Province, Thailand. To address the issue of imbalanced datasets, the model will also adapt Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) and Conditional Tabular Generative Adversarial Networks (CTGAN). The model's performance was evaluated and compared

to other existing methods in terms of accuracy, precision, recall, F-measure, and receiver operating characteristic area under the curve (ROC AUC) score. Additionally, the proposed model was compared to practical binary classification algorithms such as KNN, DT, and SVM, as well as other methods from ensemble learning groups, including RF, GB, and AdaBoost.

2.2 Theory of Computation

This section provides a fundamental exploration of the principles and capabilities of computational systems, which serve as a theoretical foundation for understanding and analyzing algorithms and computational models employed in this study.

2.2.1 Data Cleaning by Interquartile Range

Outliers can significantly impact the results of the analysis and modeling. It is crucial to ensure that our training data set does not include abnormal or significantly different data points, as these outliers can adversely affect the model's ability to learn and accurately predict outcomes. Therefore, it is essential to remove them to ensure the accuracy of the results.

The method to be selected in this study is called the interquartile range method (IQR) which is a statistical technique to measure the spread or dispersion of a dataset. The concept of calculation is based on the data points between the 1st quartile and the 3rd quartile of value as shown on the Equation (1).

$$IQR = Q_3 - Q_1 \tag{1}$$

$$Lower\ Outliers = Q_1 - (1.5 * IQR) \tag{2}$$

$$Upper\ Outliers = Q_3 + (1.5 * IQR) \tag{3}$$

Where IQR represents the Interquartile Range, with Q_1 being the 1st quartile or 25th percentile, and Q_3 being the 3rd quartile or 75th percentile. After calculating the IQR value, the data points that fall below the lower outliers (Equation 2) or above the upper

outliers (Equation 3) would be identified as outliers which are deducted before training the model.

2.2.2 One-Hot Encoding

One-Hot encoding is a widely used technique for converting nominal or categorical variables into a numeric value for machine learning models. It is especially handy when handling with discrete variables that don't possess any inherent order or numerical meaning. By applying One-Hot encoding, it ensures that the machine learning algorithms do not mistakenly assume any ordinal relationship among the categories. Instead, each category is represented by a separate binary column, enabling the models to effectively interpret the categorical information.

2.2.3 Mean Imputation

Mean imputation is a widely used method for addressing missing values in datasets by replacing them with the mean value of the respective variables. This technique is applicable across variable types, encompassing both continuous and categorical data. However, it is important to acknowledge that mean imputation can be influenced by outliers within the dataset, potentially affecting the accuracy and reliability of subsequent analyses.

2.2.4 Maximum Absolute Scaling

Maximum absolute scaling is a fundamental technique used to normalize data by dividing each observation by the maximum value of the variable. This process ensures that all values are proportionally adjusted to a uniform range relative to the variable's maximum, which promotes consistency and enhances computational effectiveness across diverse datasets. By employing maximum absolute scaling, models can improve the robustness and reliability of computational outcomes, ultimately optimizing the performance and accuracy of their processes.

2.2.5 Yeo-Johnson Transformation

Yeo-Johnson transformation is a statistical method used to normalize data and stabilize its variance. It's an extension of the Box-Cox transformation, designed to handle variables with positive, negative, or zero values. By applying a specific formula to each

data point based on its original value, Yeo-Johnson adjusts the distribution towards normality. This transformation is valuable in statistical modeling and analysis, particularly when dealing with skewed data or varying levels of variance. It helps ensure data quality and enhances the reliability of statistical inferences and model performance.

2.2.6 SMOTENC

Synthetic Minority Over-sampling Technique for Nominal and Continuous features is a method in machine learning used to handling imbalanced datasets. It builds on the original SMOTE (Synthetic Minority Over-sampling Technique) by handling datasets with both numerical (continuous) and categorical (nominal) features. This technique creates new synthetic examples for minority class instances by looking at their closest neighbors in the dataset's feature space. By doing this, SMOTENC helps balance the distribution of classes in datasets where one class is much less common than the others. This balancing act is crucial because it can improve the performance of machine learning models, particularly classifiers, by providing them with more evenly represented training data.

2.2.7 CTGAN

Conditional Generative Adversarial Network (CTGAN) is a machine learning model designed to generate synthetic data that closely resembles real-world datasets by using a generative adversarial network (GAN) framework. CTGAN learns from existing data to create new examples that replicate the statistical characteristics of the original data. While effective, CTGAN requires sufficient data to learn meaningful patterns and produce realistic synthetic datasets, making it a powerful tool for data augmentation and model training.

2.2.8 Decision Tree

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2 \tag{4}$$

$$Entropy = \sum_{i=1}^{n} - P_i log_2(P_i) \tag{5}$$

Decision Tree (DT) is a machine learning algorithm that uses a tree structure to classify subjects based on an outcome by using a splitting criterion, such as Gini impurity (Equation 4) or entropy (Equation 5) where p is a probability of each class in each node, to measure the impurity of a node. It recursively performs splitting by evaluating all possible features and thresholds to find the optimal split that maximizes information gain or reduces impurity which is allowing the algorithm to effectively partition the data and make predictions based on learned patterns. It is widely applied for prediction and classification due to its intuitive, easy understanding, high accuracy, and high prediction ability [54]. While commonly used in medical screening and diagnostics for disease prediction, decision trees are increasingly being applied in public health research to identify complex relationships between outcomes and risk factors. The insights provided by decision trees can be used to target prevention and intervention measures for the most at-risk individuals [55].

2.2.9 Random Forest

Random Forest (RF) is a powerful ensemble learning algorithm that improves accuracy by combining multiple decision trees to make predictions by randomly sampling training data sets to generate Decision Trees model and then performing aggregation prediction according to the majority voting principle to get the prediction results. It effectively handles high-dimensional data and reduces overfitting, substantially enhancing pattern recognition accuracy [58–59].

2.2.10 Extra Trees

Extra Trees algorithm, or Extremely Randomized Trees, is an ensemble learning method that builds multiple decision trees. It differs from other tree-based models by introducing additional randomness in the splitting process. Instead of searching for the optimal cut point based on the Gini index (Equation 4) or other criteria, Extra Trees randomly selects cut points. The final prediction is made by aggregating the outputs of the individual trees. Computational efficiency is a significant strength of this algorithm [69-70].

2.2.11 Gradient Boosting

$$Residual = Actual - Predicted \tag{6}$$

$$\begin{aligned} &Updated\ Model \\ &= Current\ Model \\ &+ (Learning\ Rate * Weal\ Learner\ Prediction) \end{aligned} \tag{7}$$

$$Prediction = \sum_{i=1}^{n} (Learning\ Rate * Weak\ Learner\ Predictions) \tag{8}$$

Gradient Boosting (GB) is an ensemble learning technique that utilizes multiple weak decision trees to make accurate predictions rather than using one robust model. The prediction accuracy is improved by minimizing errors by combining decision trees that reflect residuals, improving accuracy as the number of iterations increases and the residual decreases [60]. The algorithm starts with an initial model and calculates the residuals by subtracting the actual values from the model's predictions (Equation 6). Then, it literately trains weak learners on these residuals and updates the model by adding the weighted predictions of the weak learners (Equation 7), with the learning rate controlling their contribution. The prediction is obtained by combining the predictions of all the weak learners, weighted by their learning rates (Equation 8).

2.2.12 Adaptive Boosting

$$Classifier\ Weight = 0.5 * \ln\left(\frac{1 - Error}{Error}\right) \tag{9}$$

$$\begin{aligned} &Updated\ Sample\ Weight \\ &= Sample\ Weight \\ &* \exp(Classifier\ Weight * Indicator) \end{aligned} \tag{10}$$

$$Normalized\ Sample\ Weight = \frac{Sample\ Weight}{Sum\ of\ Sample\ Weight} \tag{11}$$

$$Prediction = Sign\left(\sum_{i=1}^{n}(Classifier\ Weight * Weak\ Learner\ Prediction)\right) \quad (12)$$

Adaptive Boosting (AdaBoost) is an ensemble learning algorithm that employs weighted voting to combine weak classifiers and construct strong classifiers. The algorithm aims to identify the optimal threshold in one of the data dimensions, dividing the data into two categories to distinguish between different classed based on the given features or attributes. AdaBoost starts by initializing sample weights, trains weak learners on the data, and calculates classifier weights (Equation 9) based on their performance. The algorithm then updates the sample weights (Equation 10), giving more importance to misclassified samples. To maintain balance, the sample weights are normalized (Equation 11). Finally, the weak learners' predictions are combined using their respective classifier weights through weighted voting, resulting in the final classification (Equation 12). Nevertheless, the algorithm's accuracy may need to be improved in certain circumstances, resulting in subpar generalization performance and over-fitting tendency [61].

2.2.13 K-Nearest Neighbors

$$Euclidean\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (13)$$

$$Manhattan\ Distance = |x_2 - x_1| + |y_2 - y_1| \quad (14)$$

K-Nearest Neighbors (KNN) is a data mining algorithm primarily used for classification tasks. It determines the class of an instance based on the majority class of its k nearest neighbors in the feature space. KNN relies on a local neighborhood approach, where it considers a few adjacent instances to make the classification decision, making it particularly suitable for datasets with overlapping classes. Unlike some algorithms, KNN does not explicitly identify the class boundaries or domain; instead, it focuses on the nearby instances for classification [53]. It starts by calculating the distances between the instance to be classified and all instances in the training set which need to specify a hyperparameter. Then, it selects the k nearest neighbors based on these distances using distance metrics such as Euclidean or Manhattan distance. Next, it determines the majority

class among these neighbors for classification. Finally, it assigns the instance to the determined class. KNN's reliance on local neighbors and its straightforward nature makes it a versatile algorithm, particularly suitable for classification tasks involving overlapping classes.

2.2.14 Support Vector Machine

Support Vector Machine (SVM) is a powerful algorithm widely utilized for classification tasks, known for its excellent performance [56]. It is a flexible and effective classifier, particularly beneficial for studies with limited sample sizes. The SVM algorithm involves five key steps. Firstly, the data is prepared, ensuring labeled training data is available. If necessary, feature transformation techniques such as the kernel trick can be applied. Next, the model is trained by finding an optimal hyperplane that maximizes the margin between classes using an optimization problem. SVM has found applications in various domains, including the diagnosis and prognosis of brain diseases like Alzheimer's disease, schizophrenia, and depression [57].

2.2.15 Pearson Correlation Coefficient

The Pearson method was used to visualize the correlation coefficient between the features. The Pearson correlation coefficient (r) is a widely used measure of linear correlation between two variables, ranging from -1 to 1. It quantifies the strength and direction of the relationship between the variables, with a value of -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation, as in Equation (15). The interpretation is depicted in Table 2.1.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{15}$$

Where r is correlation coefficient, x_i is value of x-variable in the dataset, $\bar{x}$ is mean of x-variable in the dataset, y_i is value of y-variable in the dataset, and $\bar{y}$ is mean of y-variable in the dataset.

Table 2.1 Pearson Correlation Coefficient Interpretation

Pearson Correlation Coefficient	Correlation Type	Interpretation
Between 0 and 1	Positive Correlation	The feature change consistency in the same direction.
0	No Correlation	There is no correlation between the features.
Between 0 and -1	Negative Correlation	The features change consistency in the opposite direction

2.2.16 Shapley Value

SHAP (SHapley Additive exPlanations) values are a method used in machine learning to explain individual predictions by attributing the contribution of each feature to the model's output. SHAP values quantify the impact of features on predictions by considering all possible combinations and their contributions. They provide interpretable insights into how each feature influences predictions, including its direction and magnitude of effect. This capability makes SHAP values valuable for understanding feature importance, interactions between features, and overall model behavior across different types of machine learning models.

2.2.17 Evaluation Metrics

Evaluation metrics are quantitative measures used to evaluate the performance and effectiveness of a machine learning model. These metrics can indicate of the model's predictive capabilities and help in comparing the different models to optimize the model's performance. In this study, the research utilized five metrics to evaluate the model performance:

1. Accuracy is a ratio of correctly predictive results of observations to total observations, as shown in Equation (16).

(16)

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

2. Precision is the proportion of correctly predicted positive observations to all predicted positive observations, as shown in Equation (17).

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

3. Recall or Sensitivity is the proportion of correctly predicted positive observations to all observations in the positive of the actual class as in Equation (18).

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

4. F1 Score is the weighted average calculation of the precision and recall value as in Equation (10), to provides a balanced measure of the model's performance.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{19}$$

5. AUC-ROC Score measures the separability and performance of a classification model. It provides a single value to quantify the capability of distinguishing between positive and negative instances, with higher values indicating better performance. It's an iterative calculation the value of True Positive Rate (Recall) and False Positive Rate (FPR, as shown in Equation 20). The final formula shows in Equation (21).

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

$$AUCROC = \sum_{i=0}^{n} (TRP[i] * (FPR[i+1] - FPR[i])) \tag{21}$$

CHAPTER 3

RESEARCH METHODOLOGY

The research methodology of the study is illustrated in Figure 3.1, covering data collection, data preprocessing, feature engineering, and model construction.

3.1 Research Overview

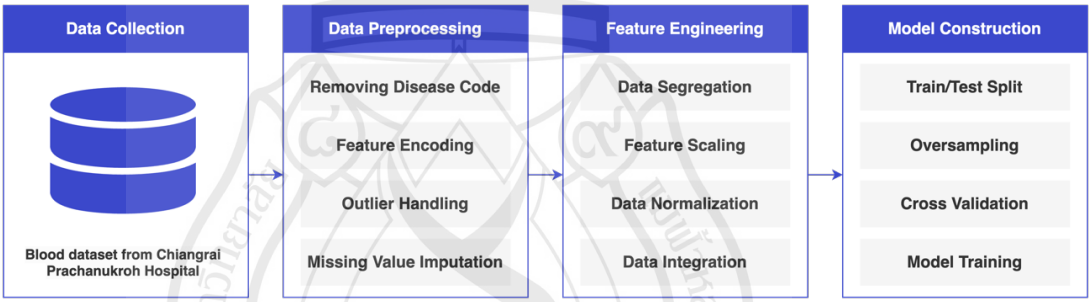

Figure 3.1 Overview of the Methodology Process Diagram

As shown in Figure 3.1, this study aims to classify diseases that may pose a risk to patients by utilizing machine learning techniques. The objective is to develop a classification model that accurately identifies a risk disease based on patient-related factors. To achieve these objectives, this study employed a 4-step process in this study, data collection, data preprocessing, feature engineering, and model construction. The study received ethical approval from the Mae Fah Luang University Ethics Committee, as documented in Appendix C, which ensured that all research procedures adhered to the institution's ethical standards, particularly in the use of patient-related data.

3.2 Data Collection

Features	Description	Type	Dementia (n = 658)	Heart Failure (n = 3,639)	P-Value
Age	Age is an essential factor when considering phenotypic changes in health and disease	Integer	75.08 ± 8.31	73.73 ± 8.73	0.0001
Biological Characteristics	Physical differences between males and females	Category	Male = 265 Female = 393	Male = 1,595 Female = 2,044	0.0901
Creatinine	A waste product resulting from the deterioration of muscles	Float	1.17 ± 0.98	2.23 ± 2.34	3.8690E-74
Blood Urea Nitrogen	A waste product removed from the blood by the kidney	Integer	18.13 ± 10.26	34.28 ± 25.68	9.6746E-110
Hemoglobin	Proteins in red blood cells from carrying oxygen to the organs and tissues of the body	Float	11.85 ± 1.82	11.02 ± 2.47	5.2481E-20
Potassium	A chemical element that helps maintain normal levels of fluid inside cells	Float	3.89 ± 0.54	3.96 ± 0.74	0.0159
White Blood Cells	A type of blood cell made in the bone marrow and found in the blood and lymph tissue	Integer	7,365.80 ± 2,864.80	9,287.45 ± 4,700.07	2.6933E-36
Neutrophils	The immune system uses neutrophils to combat infections and speed up wound healing	Float	60.22 ± 10.76	71.50 ± 14.30	5.5805E-83
Platelets	Tiny, colorless cell fragments that help stop or slow bleeding	Integer	255,473.20 ± 88,147.84	222,021.65 ± 109,599.14	8.1102E-15
Lymphocytes	An immune system component that is a type of white blood cell	Float	26.50 ± 8.79	18.05 ± 10.76	2.2791E-73

Figure 3.2 Data Attributes

Figure 3.2 presents the data collected from Chiang Rai Prachanukroh Hospital, Chiang Rai Province, Thailand, comprising 4,297 records. The dataset includes 10 blood test features categorized into dementia (represented by the codes F00, F01, F02, and F03) and heart failure (represented by the code I50) based on the International Classification of Diseases 10th Revision (ICD-10) codes.

Exploratory Data Analysis (EDA) was conducted to further examine these findings and provide graphical representations of the distribution and correlation of features in the dataset. These visual representations (Appendix A) significantly contributed to understanding their importance within the context of the study. Based on the visualizations in Appendix A, biological characteristics and hemoglobin emerged as significant features during the analysis. The dataset revealed a higher incidence of both conditions among females, which is consistent with established statistics indicating that females are more susceptible to dementia [71–73] and heart failure [73–74]. Additionally, hemoglobin levels, recognized as a potential biomarker for disease activity [42–43], showed relevance to both heart failure and dementia risk [75–76].

Furthermore, a comprehensive statistical analysis was conducted using a two-tailed distribution and two-sample unequal variance testing, which revealed low p-values associated with key factors: creatinine, blood urea nitrogen, hemoglobin,

white blood cells, neutrophils, platelets, and lymphocytes. These extremely low p-values indicated strong statistical significance. Typically, a p-value of 0.05 or lower is considered the standard threshold for determining statistical significance. However, the values obtained for these factors were notably smaller, suggesting an exceedingly low probability of these results occurring by chance. Specifically, a p-value close to zero suggests that the observed relationships between these blood test features, and the targeted outcomes (dementia and heart failure) were highly unlikely to be random fluctuations in the data. Such insights underscore the predictive potential and consequential impact of these features.

3.3 Data Preprocessing

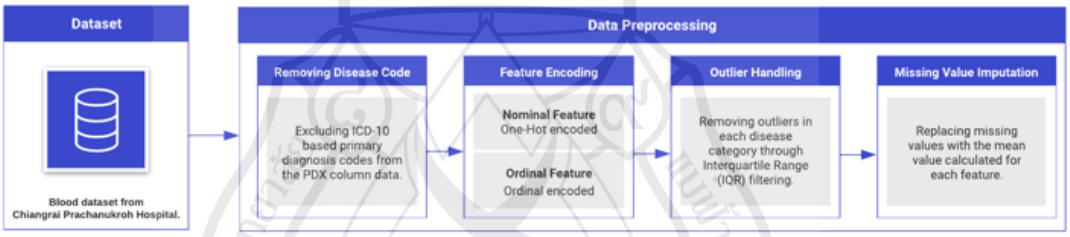

Figure 3.3 Data Preprocessing Workflow Diagram

As shown in Figure 3.3, data preprocessing is a critical initial phase that refines raw data to enhance completeness and enable efficient training for subsequent analysis and modeling. This process ensures the quality of the data and sets the groundwork for further analysis.

3.3.1 Removing the Disease Code

The specific code forms of the principal diagnosis (PDX) variables associated with disease-indicative ICD-10 codes were excluded from this study. These codes directly indicate the patient's illnesses, which could potentially give the classifier an unfair advantage. This exclusion was done while keeping other disease codes that are not related to dementia or heart failure.

3.3.2 Feature Encoding

To enhance the learning process of machine learning algorithms, this study employed specific encoding techniques to transform the data. Nominal data was subjected to one-hot encoding, while categorical data was processed using ordinal encoding. These transformations allowed the algorithm to effectively interpret and learn from the dataset.

3.3.3 Outlier Handling

The principle of interquartile range, as described in Equation (1), was used in this study to identify and remove outliers from the dataset. Removing outliers was considered essential due to their potential to greatly impact analysis and modeling results. Specifically, this study examined data that fell within the 25th and 75th percentiles.

3.3.4 Missing Value Imputation

To ensure effective learning within the model, it was considered crucial to handle the missing data in the dataset. In this study, the missing values were replaced with the mean value of the feature. This approach was chosen to enable the model to learn from a complete dataset and thereby improve its performance.

3.4 Feature Engineering

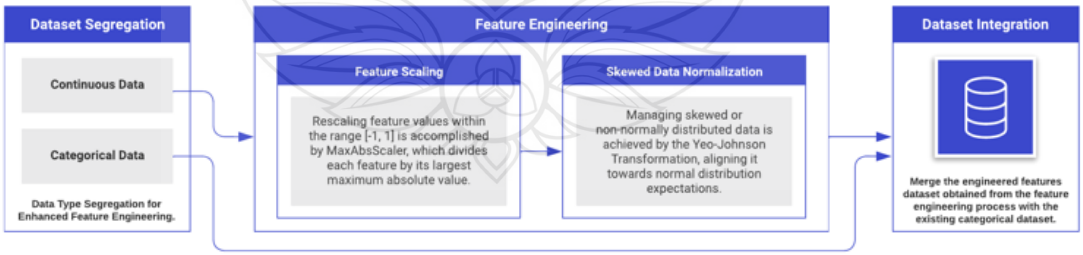

Figure 3.4 Feature Engineering Process Diagram

As shown in Figure 3.4, after receiving the preprocessed data, the feature engineering process began by separating the data into categorical and continuous types. Continuous features were then subjected to a maximum absolute scaler and Yeo–Johnson transformation. Once optimized, these features were combined for model construction.

3.4.1 Dataset Segregation

After preprocessing the dataset, it was divided into two distinct categories: continuous and categorical. Subsequently, this study focused on enhancing the continuous data through feature engineering. The goal was to enrich the dataset with more informative attributes, thereby improving machine learning capabilities for more accurate predictive outcomes.

3.4.2 Feature Scaling

The maximum absolute scaler used in this study to normalize the data offered several advantages. By adjusting the scale of each feature relative to its maximum absolute value, this technique ensured that all features were uniformly represented within a range of –1 to 1. This normalization not only standardizes the data but also reduces the influence of outliers, making the dataset more robust against extreme values. Moreover, this process preserves the relative relationships between feature values while enabling fair and accurate comparisons across features. This improvement enhances the interpretability and performance of the subsequent analysis and modeling.

3.4.3 Skewed Data Normalization

The Yeo–Johnson transformation was used in this study to address data skewness, improving model accuracy and ensuring more reliable predictions. Skewed data can significantly hinder machine learning model performance because many algorithms assume a normal distribution. Excessive distortion in the data can lead the model to learn inaccurate patterns, misrepresent true data relationships, and introduce bias into predictions. Rectifying skewed data not only improves interpretability by establishing clearer variable relationships but also mitigates these issues.

3.4.4 Dataset Integration

After performing feature engineering on the continuous data, the refined dataset was combined with the previously separated categorical data. This merger was done as a preparatory step to facilitate dataset partitioning for the subsequent model construction.

3.5 Model Construction

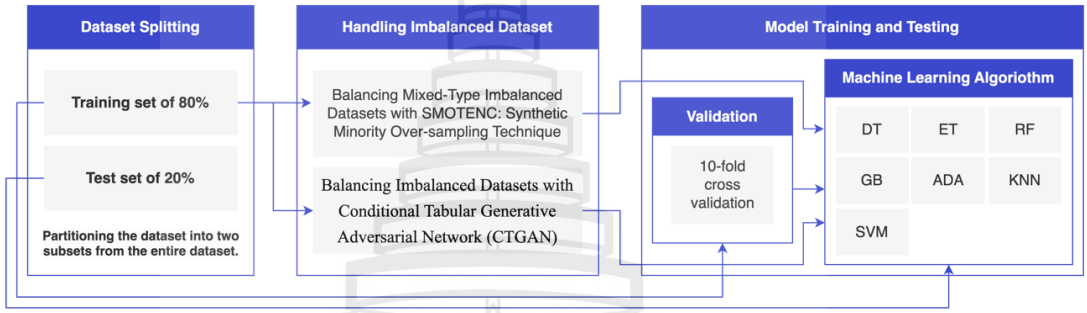

Figure 3.5 Overview of the Model Construction Process

As shown in Figure 3.5, the model construction process began with dividing the initial data into training and test sets. After that, the training dataset underwent 10-fold cross-validation to validate the ET model [77]. Subsequently, the ET classifier was trained using the oversampled dataset from the training set, along with the training of other classifiers, to evaluate their performances.

3.5.1 Dataset Splitting for Training and Testing

As recommended in [78], the dataset was intentionally divided into an 80% training dataset and a separate 20% test dataset. This strategic division facilitated fair model training and evaluation. Such a deliberate approach was chosen to ensure unbiased representation in both subsets, which is a crucial factor for robust model development.

3.5.2 Oversampling Technique

3.5.2.1 Handling Mixed-Type Imbalanced Data Using SMOTENC

Due to the significant disparity in the ratio between dementia and heart failure data, this study utilized the synthetic minority oversampling technique for nominal and continuous features (SMOTENC) to address this imbalance. This technique specifically focuses on augmenting the training data, without affecting the testing data, thereby enhancing the reliability of the model by rectifying the data imbalance. This deliberate approach ensures a fairer representation within the training set, which is crucial for achieving robust model performance when dealing with imbalanced datasets.

3.5.2.2 Generating Balanced Datasets with CTGAN

In addition to SMOTENC, this study employs Conditional Tabular Generative Adversarial Networks (CTGAN) to address imbalances in datasets with mixed types of data. CTGAN generates synthetic data that closely mimics real data, preserving complex relationships across different types of features (e.g., categorical and numerical). This approach helps balance the dataset by creating new samples for underrepresented classes. By maintaining the dataset's original distribution, CTGAN enhances model training, leading to more accurate disease classification.

3.5.2.3 Model Training and Testing

In this study, ETs were used to build the model. ETs use multiple DTs and combine their predictions from the blue nodes of each decision tree to improve accuracy, as shown in Figure 3.6. Unlike other tree-based models, ETs have a different approach to splitting. Instead of exhaustively searching for the best split at each node, they randomly select cut points from the original training samples for each feature and generate trees without bootstrap data. This maintains computational efficiency while increasing the diversity of the trees within the ensemble. This randomness helps to reduce the correlation between individual trees in the ensemble, which reduces overfitting and improves generalization performance. In this study, the Gini index is used as the criterion to evaluate the quality of splits, ensuring the effectiveness of the model’s decision-making process [69–70]. To ensure the classification performance of the ET-based model being proposed, it is essential to compare it with other machine learning-based methods. Specifically, the proposed ET model is compared to practical

binary classification algorithms such as KNN, DT, and SVM, as well as other methods from the ensemble learning groups, including RF, GB, and AdaBoost. For the validation process, this study employed 10-fold validation for all the constructed models.

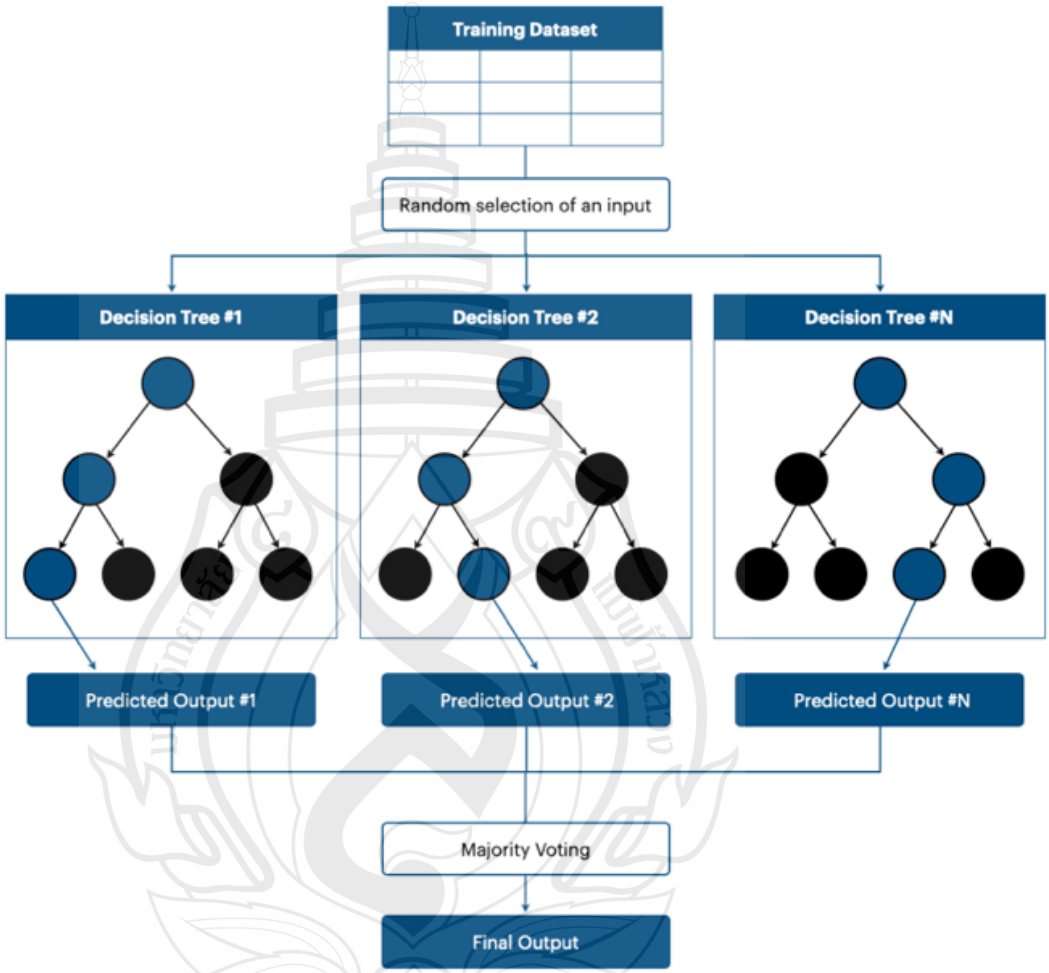

Figure 3.6 Extra Trees Classifier Diagram

3.6 Model Evaluation

Table 3.1 Prediction Metrics for Assessing Model Performance

Class		**Predicted Class**	
Actual Class	Disease	Heart Failure	Dementia
	Heart Failure	True Positive (TP)	False Negative (FN)
	Dementia	False Positive (FP)	True Negative (TN)

The performance of the machine learning classifier is assessed using various metrics, such as accuracy, recall, precision, F-measure, and ROC AUC score. Four variables are used to determine the evaluation parameters: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In this study, TP and FN indicate heart failure, while FP and TN represent dementia, as shown in Table 3.1.

3.7 Feature Evaluation

Feature evaluation is a crucial process for understanding the influence of individual features within a model. This assessment typically involves two fundamental methods: the Pearson correlation coefficient, which measures feature correlations, and the SHapley Additive exPlanation (SHAP) value, which reveals the impact of each feature on the target variable.

3.7.1 Pearson Correlation Coefficient

The Pearson correlation coefficient, a frequently used measure of linear correlation that ranges from –1 to 1 between two variables, was used to visualize the correlation coefficient between features and evaluate their correlation before implementation. It measures the strength and direction of the relationship between the variables, with a value of –1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation, as shown in Equation [15]. The interpretation of the coefficient values is provided in Table 2.1.

3.7.2 Feature Importance by Using Extra Trees

This study utilizes Extra Trees (ET) to assess feature importance to understand the impact of insights into the importance of features for model predictions. ET, an ensemble learning technique similar to Random Forests but with extra randomness in the node splitting, evaluates the effectiveness of each feature in reducing impurity across multiple decision trees. By aggregating these assessments, ET ranks features based on their contribution to model accuracy, thereby aiding in the identification of critical biomarkers or variables associated with disease classification. This method enhances the interpretability of the classifier model by highlighting the most influential features driving predictions.

3.7.3 Shapley Value

To gain a thorough understanding of how different variables affect model predictions, effective methodological approaches are essential. Among the various techniques available, priority is given to SHAP values, widely recognized and extensively utilized. SHAP values offer several advantages in enhancing model interpretability, not only by providing insights into variable importance but also by revealing their directional impact on predictions [79]. Additionally, they allow for a detailed understanding of the contributions of individual features to specific predictions. Importantly, the adaptability of SHAP values across diverse models distinguishes them from methods limited to specific model types, making SHAP a versatile and inclusive tool for comprehensive model analysis. The SHAP technique encapsulates the essence of Shapley values, capturing the average marginal contribution of each feature to overall predictions and highlighting their significance in various feature combinations. In this study, SHAP is employed to evaluate feature contributions using test data, facilitating a detailed examination of individual feature impacts on the model's predictions.

CHAPTER 4

EXPERIMENTAL RESULTS

This section presents the results obtained from using seven different machine learning classifiers on the datasets. To ensure reliability and credibility, this study first demonstrate the outcomes of 10-fold cross-validation, which emphasizes the trustworthiness of the models. Then, it employs important evaluation metrics, including accuracy, precision, recall, F-measure, and ROC AUC score, to evaluate the performance of these classifiers.

4.1 Pearson Correlation Coefficient

Based on the Pearson correlation coefficient presented in Figure 4.1, the analysis results have identified significant positive correlations between the targeted class and lymphocytes (27%), hemoglobin (12%), platelets (11%), and age (6%). Conversely, there is a negative correlation with neutrophils (–27%), blood urea nitrogen (–20%), creatinine (–17%), and white blood cells (–15%).

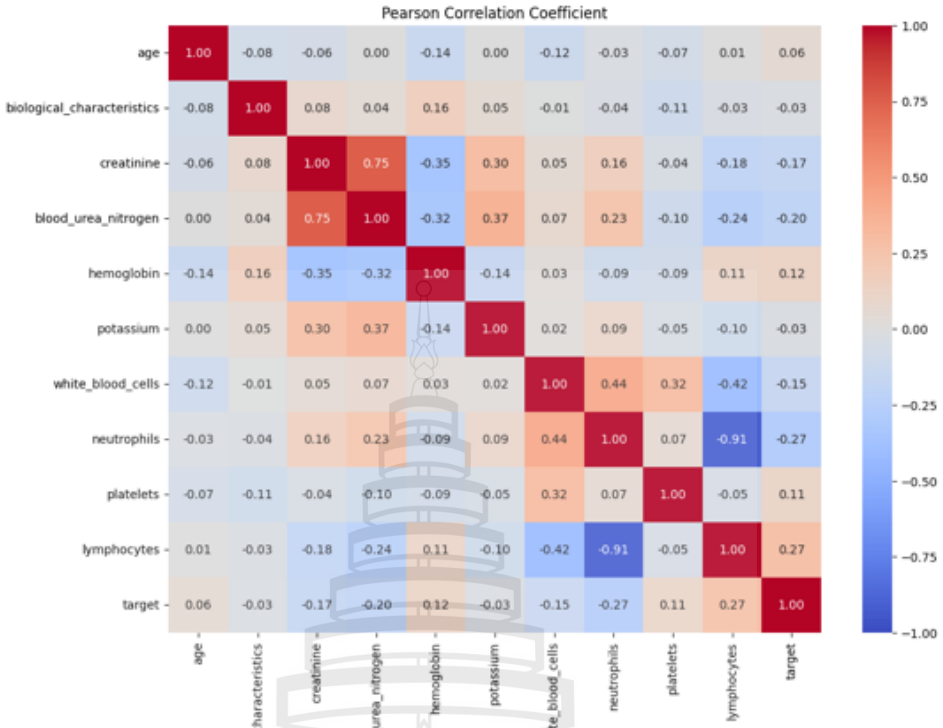

Figure 4.1 Feature Correlation Measurement by Pearson Correlation Coefficient

In the analysis of the correlation between the study variables, several significant findings have emerged. Firstly, hemoglobin and platelets, both relevant in disease activity, exhibited a negative correlation of 9%. This suggests that as hemoglobin levels increase, platelet counts tend to decrease. Secondly, the association between white blood cells and their constituents, neutrophils and lymphocytes, revealed interesting patterns. There was a positive correlation of 44% between white blood cells and neutrophils, indicating that as white blood cell counts increase, neutrophil levels also tend to increase. Conversely, there was a negative correlation of –42% between white blood cells and lymphocytes, implying an association between increased white blood cell counts and a decrease in lymphocyte levels. Furthermore, the highest correlation strength in the dataset was observed between neutrophils and lymphocytes, with a negative correlation of –91%, indicating a strong linkage between an increase in neutrophil levels and a decrease in lymphocyte levels. Another notable finding was the second-highest correlation of 75% between creatinine and blood urea nitrogen, both used to assess kidney function. This positive correlation suggests

that as creatinine levels increase, blood urea nitrogen levels also tend to increase. Finally, concerning their correlation with the target variable, certain features exhibited notable associations: lymphocyte (27%), neutrophil (–27%), blood urea nitrogen (–20%), and creatinine (–17%). These percentages delineate both positive and negative correlations observed in this analysis, illustrating the degrees of association to effectively incorporate into the model.

Furthermore, Principal Component Analysis (PCA) was used to reduce the dataset's dimensionality by focusing on fewer components. However, this approach did not improve key performance metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. This indicated that the reduced components didn't capture enough of the data's variance to maintain strong model performance. As a result, it became clear that retaining all features was necessary to preserve the important relationships in the data and achieve the best possible model predictions. A comprehensive analysis of the performance metrics resulting from the adaptation of PCA is presented in Appendix B.

4.2 Cross-Validation Evaluation

Table 4.1 Classification Performance Metrics on Cross-Validation

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	87.27%	92.88%	92.45%	92.66%	72.72%
GB	85.10%	94.94%	87.54%	91.07%	78.27%
Ada	81.59%	95.24%	82.97%	88.66%	77.73%
RF	87.53%	93.56%	91.99%	92.76%	75.00%
DT	80.91%	92.04%	85.42%	88.59%	68.21%
KNN	73.85%	94.54%	74.21%	83.13%	72.85%
SVM	72.01%	96.53%	70.32%	81.34%	76.7%

Table 4.1 illustrates the use of 10-fold cross-validation to improve the reliability of model performance assessment. The average scores obtained across iterations were as follows: RF (87.53%), ET (87.27%), GB (85.10%), AdaBoost (81.59%), DT (80.91%),

KNN (73.85%), and SVM (72.01%). These scores, obtained by dividing the dataset into ten subsets for repeated evaluation, reflect the models' consistent predictive abilities. Notably, RF and ET showed strong performances, demonstrating high reliability and consistency. GB and AdaBoost also maintained solid averages of 85.10% and 81.59%, respectively. However, DT, KNN, and SVM showed comparatively lower average scores, suggesting potential variability or limitations in their predictive capacities across the cross-validated datasets. The use of 10-fold cross-validation provided a comprehensive and reliable evaluation of the models' performance.

4.3 Performance Evaluation Metrics for Classification Models

During the testing process, performance metrics such as accuracy, precision, recall, F-measure, and ROC AUC score were used for comparison. Detailed comparison results can be found in Table 4.2.

Table 4.2 Classification Performance Metrics with SMOTENC

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	89.11%	94.92%	92.50%	93.70%	78.88%
GB	85.04%	94.81%	87.71%	91.12%	77.01%
Ada	83.46%	96.24%	84.41%	89.94%	80.62%
RF	86.61%	93.66%	90.85%	92.24%	73.85%
DT	83.07%	93.25%	86.96%	89.99%	71.37%
KNN	76.51%	95.02%	77.21%	85.19%	74.40%
SVM	73.36%	96.22%	72.41%	82.63%	76.21%

4.3.1 Accuracy Evaluation

According to the analysis of the implemented prediction models, ETs stood out by outperforming all other models with an impressive accuracy of 89.11%. The accuracy, which represents the proportion of correct predictions, highlighted ET's capability to accurately classify data points within the evaluated dataset. Its superior performance

compared to RF (86.61%), GB (85.04%), AdaBoost (83.46%), DT (83.07%), KNN (76.51%), and SVM (73.36%) underscored its effectiveness in making precise predictions. The utilization of a random feature subset selection and node splitting strategy within the ensemble learning framework likely contributed to ET's reduced variance. This approach mitigates overfitting concerns, ultimately resulting in the highest reliability and accuracy among the models assessed for this specific prediction task.

4.3.2 Precision Evaluation

Precision, a critical metric for classification models, represents the accuracy of positive predictions within the predicted class. In this study, AdaBoost, SVM, and KNN demonstrated exceptional precision scores of 96.24%, 96.22%, and 95.02%, respectively, excelling in identifying positive instances while minimizing FPs. On the other hand, models like ET, GB, RF, and DT, despite exhibiting higher accuracy, presented comparatively lower precision scores, indicating a trade-off between overall accuracy and precision. Models that achieve higher accuracy may capture a larger number of instances, but they tend to lack precision when distinguishing specific classes, resulting in an increase in FPs. However, high-precision models prioritize minimizing FPs, which may lead to the potential omission of some TPs and impact overall accuracy. Despite ET's slightly lower precision score, its ensemble method, which utilizes multiple decision trees and randomness in feature selection, ensures a robust overall predictive capability, making it a formidable choice for predictive modeling across diverse datasets and noisy environments.

4.3.3 Recall Evaluation

Recall, a metric used to evaluate a model's ability to predict the correct class output, quantifies the ratio of TP predictions to the total number of instances belonging to the actual class in the dataset. Among the models examined in this study, ETs (92.50%), RF (90.85%), GB (87.71%), and DT (86.96%) emerged as the top performers in terms of recall. Notably, these models demonstrated strong capabilities in accurately identifying instances of the actual class in the dataset. In contrast, AdaBoost (84.41%), KNN (77.21%), and SVM (72.41%) achieved comparatively lower recall scores. This variation in recall scores highlights the differences in the effectiveness of these models in capturing true positives within their predicted classes. However, the models with higher recall scores

significantly enhanced their overall predictive modeling performance, emphasizing their reliability in handling classification tasks.

4.3.4 F1 Score Evaluation

The F1, a comprehensive evaluation metric derived from the harmonic mean of precision and recall, serves as a single measure of a model's effectiveness, eliminating the need to prioritize between precision and recall. Among the evaluated models, ETs (93.70%), RF (92.24%), GB (91.12%), and DT (89.99%) emerged as the top performers in terms of F-measure. These models displayed a balanced combination of precision and recall, indicating their robustness in accurately identifying the predicted class while considering the actual class instances. On the other hand, AdaBoost (89.94%), KNN (85.19%), and SVM (82.63%) demonstrated relatively lower F-measure scores, indicating potential variations in their trade-offs between precision and recall. Specifically, ET and RF stood out in terms of the F-measure, emphasizing their strength in achieving a balance between precision and recall, thereby solidifying their expertise in predictive modeling tasks.

4.3.5 AUC-ROC Score Evaluation

The ROC AUC, a metric used to evaluate the performance of binary classification models, measures their ability to distinguish between different classes. In this study, AdaBoost (80.62%), ETs (78.88%), GB (77.01%), and RF (73.85%) demonstrated the highest ROC AUC scores in distinguishing between dementia and heart disease. Importantly, these models showed strong discriminatory power in accurately classifying the two classes. SVM (76.21%) and KNN (74.40%) also performed reasonably well in this context. However, DT lagged behind, with an ROC AUC score of 71.31%, suggesting its comparative inefficiency in this specific task. The results highlight the effectiveness of ensemble-based methods, particularly AdaBoost, ET, GB, and RF, in effectively addressing this binary classification challenge, demonstrating their potential for distinguishing between dementia and heart disease.

4.3.6 Synthetic Minority Oversampling Technique for Nominal and Continuous Features Evaluation

SMOTENC, a variation of the synthetic minority oversampling technique (SMOTE), is specially designed to handle imbalanced datasets with both numerical and categorical features. In this study, this study applied SMOTENC preprocessing to address the class imbalance present in the dataset. When compared to the proposed model, models incorporating SMOTENC showed higher ROC AUC scores than those without SMOTENC, indicating improved performance in detecting instances of the minority class, especially dementia cases. Although the method without SMOTENC demonstrated higher accuracy rates, their lower ROC AUC scores suggest a tendency for bias towards the majority class, which hinders effective detection of instances in the minority class.

A high recall score indicates the ability to accurately identify positive instances, which may result in a higher number of false positives due to an imbalance in classes throughout the dataset. Consequently, the precision score tends to decrease without SMOTENC, indicating false positives in all instances classified as positive. The application of SMOTENC not only helps improve the ROC AUC score but also enhances overall predictions by reducing errors from false positives, particularly in situations where class imbalance is prevalent. Although the model without SMOTENC has higher accuracy, the inclusion of SMOTENC can effectively address the imbalance problem, resulting in improved ROC AUC scores and reduced false positive errors.

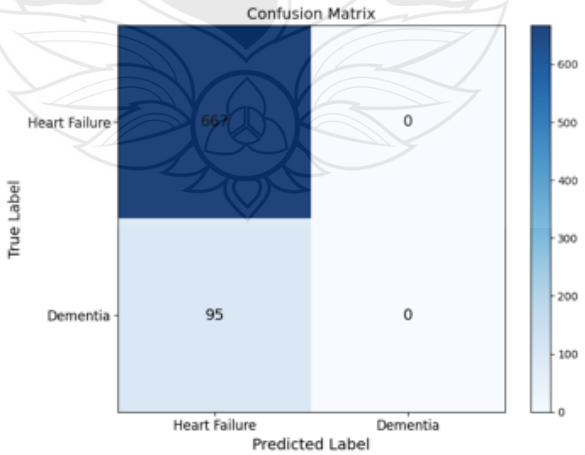

Figure 4.2 Confusion Matrix for SVM Algorithm Trained without SMOTENC

Table 4.3 Classification Performance Metrics without SMOTENC

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	90.81%	91.06%	99.25%	94.98%	65.41%
GB	89.76%	91.89%	96.85%	94.31%	68.43%
Ada	89.24%	91.61%	96.55%	94.01%	67.22%
RF	90.03%	91.56%	97.60%	94.48%	67.22%
DT	83.07%	90.76%	89.81%	90.28%	62.80%
KNN	88.19%	91.27%	95.65%	93.41%	65.72%
SVM	87.53%	87.53%	100.00%	93.35%	50.00%

As shown in Table 4.3, the SVM results without SMOTENC demonstrate a 100% recall rate, indicating the correct identification of all positive instances. However, Figure 4.2, which displays the confusion matrix of SVM after training without SMOTENC, reveals a critical drawback: the model completely ignores negative instances, resulting in a lack of true negatives and overall substandard performance. This highlights the need for SMOTENC preprocessing to address such limitations. Therefore, emphasizing the use of SMOTENC-preprocessed models becomes an important observation to enhance prediction capabilities, especially in critical applications such as medical diagnosis, where accurate identification of minority instances is crucial.

4.3.7 Conditional Tabular Generative Adversarial Network Evaluation

Table 4.4 Classification Performance Metrics with CTGAN

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	85.96%	92.55%	91.30%	91.92%	69.86%
GB	86.35%	92.72%	91.60%	92.16%	70.54%
Ada	83.73%	93.30%	87.71%	90.42%	71.75%
RF	85.17%	93.42%	89.36%	91.34%	72.57%
DT	83.86%	93.73%	87.41%	90.46%	73.18%
KNN	85.30%	92.37%	90.70%	91.53%	69.04%
SVM	80.97%	92.23%	85.46%	88.72%	67.47%

According to Table 4.4, the evaluation shows that while the top two classifiers using SMOTENC, ET and RF, have lower performance with CTGAN, the remaining classifiers (GB, AdaBoost, DT, KNN, and SVM) perform better in terms of accuracy. However, focusing on AUC-ROC and precision scores reveals a significant drop with CTGAN across nearly all classifiers, indicating that CTGAN cannot effectively help models predict the minority class compared to SMOTENC. This is likely due to CTGAN's less effective representation of the minority class in the generated synthetic samples, which impacts the models' ability to distinguish between classes accurately.

This result can be attributed to the fundamental differences in the data generation approaches of SMOTENC and CTGAN. SMOTENC generates new samples by interpolating between existing minority class samples using nearest neighbors, which ensures that the synthetic data is close to actual minority instances and directly addresses the class imbalance. This targeted generation enhances the model's ability to learn the characteristics of the minority class effectively. Conversely, even when CTGAN is used to generate synthetic data solely for the minority class, it creates an entire dataset based on the learned distribution. Although CTGAN can produce more diverse and realistic data overall, its focus on generating a broad distribution rather than specifically interpolating minority class instances may result in less precise representation. This can lead to poorer performance in distinguishing between classes, as reflected by the lower ROC AUC scores when using CTGAN.

4.3.8 Feature Importance Using ET

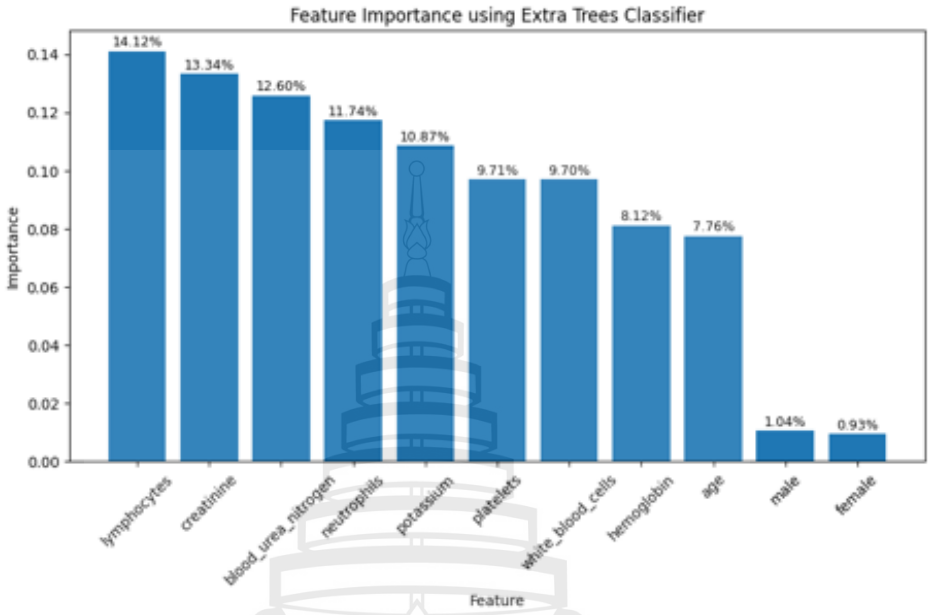

Figure 4.3 Feature Importance Using the Extra Trees Classifier

After implementing the ET model, this study conducted a comprehensive analysis of the key factors that influence prediction outcomes. As shown in Figure 4.3, the analysis revealed that the most influential factors were lymphocytes (14.12%), creatinine (13.34%), blood urea nitrogen (12.60%), neutrophils (11.74%), potassium (10.87%), platelet count (9.71%), white blood cells (9.70%), hemoglobin (8.12%), and age (7.76%). These factors had a significant impact on the model's predictions, highlighting their important role in determining outcomes. However, it is important to note that certain features, such as indicators associated with biological characteristics (male: 1.04%, female: 0.93%) had minimal impact. Their coefficients, measuring below 2%, indicated that they had a negligible influence on the predictive accuracy of the model. Although these features were included in the analysis, they were found to have little relevance to the overall predictions.

4.3.9 Impact of Features on Model Prediction

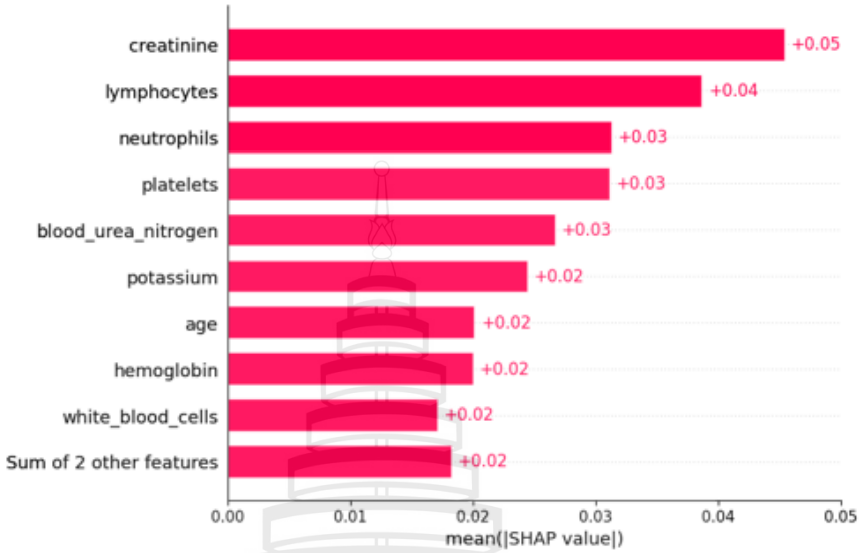

Figure 4.4 Mean Shapley Values for Feature Impact

This study examines the comprehensibility of a predictive model using SHAP values calculated from a test dataset. As shown in Figure 4.4, the average SHAP values clarify the influence of features on the predictions generated by the best model suggested using the ET model. Specifically, creatinine emerges as the most influential feature, exhibiting a significant positive effect (+0.05), closely followed by lymphocytes (+0.04), neutrophils (+0.03), platelets (+0.03), and blood urea nitrogen (+0.03). Moreover, features such as potassium, age, hemoglobin, and white blood cells display slightly lower yet consistently positive impacts (+0.02). Remarkably, the combined consideration of biological characteristics features also contributes significantly (+0.02 on average). These findings highlight the varying contributions of features to the model's predictive outcomes, emphasizing the heightened importance of creatinine, lymphocytes, and neutrophils in influencing the model's predictions. This observation aligns with the correlations revealed in the heatmap generated by the Pearson Correlation Coefficient, confirming their pivotal role in assessing feature relevance for model implementation. Furthermore, it aligns seamlessly with the results obtained from feature importance analysis. This consistency

across different methodologies strengthens confidence in the utilization of these key features within the classification model.

4.3.10 Confusion Matrix Evaluation

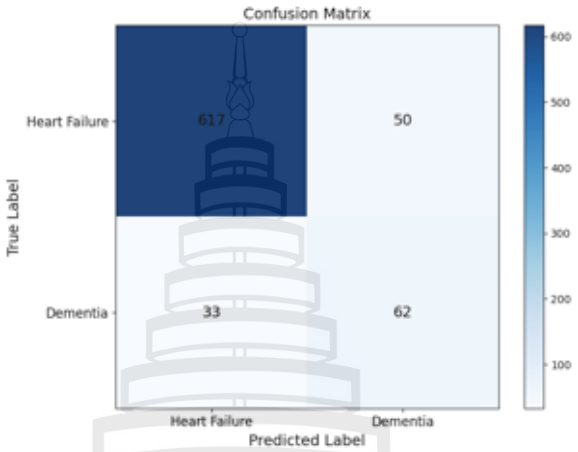

Figure 4.5 Confusion Matrix using the Extra Trees Classifier

From the test results, Figure 4.5 illustrates the confusion matrix, which showcases the classification performance of the proposed ET model in distinguishing between the two classes: heart failure and dementia. The number of accurate TPs indicated heart failure with a count of 617. Conversely, FNs indicated instances where the model incorrectly predicted dementia instead of heart failure, totaling 50 instances. FPs represented instances where the model mistakenly predicted heart failure instead of dementia, with a count of 33. Lastly, TNs represented accurate predictions for the dementia class, with a count of 62.

4.4 In-Depth Analysis of Oversampling Techniques

SMOTENC is a technique used to address class imbalance in datasets that include both categorical and continuous features. It operates by identifying the nearest neighbors for each minority class sample and generating synthetic samples through interpolation between these neighbors. This approach ensures that the new samples are closely related to the existing data, preserving the underlying patterns and class distribution. By focusing on

the nearest neighbors, SMOTENC creates realistic and meaningful synthetic samples that enhance the minority class without distorting the dataset's structure.

On the other hand, CTGAN is a type of Generative Adversarial Network specifically used to generate synthetic data. CTGAN involves two neural networks: a generator that creates new data samples and a discriminator that tries to distinguish between real and fake samples. By learning from the existing data, the generator is conditioned on class labels to produce samples that are representative of the minority class. While CTGAN can generate a wide variety of samples, its adversarial nature may lead to the creation of less realistic or noisy data, especially when the original dataset is small. This variability can sometimes make CTGAN less reliable for datasets with limited examples.

Given the limitations associated with dataset instance constraints, SMOTENC is often preferred due to its ability to generate synthetic samples through the use of nearest neighbors, which effectively reduces the risk of introducing noise. While CTGAN is proficient at modeling complex data patterns, it may produce less realistic samples, especially when working with smaller datasets. As a result, SMOTENC is generally considered more reliable for maintaining data quality.

4.5 Discussion of Overall Disease Prediction

The analysis presented in Table 4.2 highlights significant differences among classifiers derived from ensemble learning methods when compared to traditional supervised learning approaches. Across various metrics, the ensemble methods consistently outperformed their traditional counterparts, with the ET model emerging as the standout performer. ET demonstrated superiority among the evaluated models for disease prediction, boasting an impressive accuracy of 89.11% alongside robust precision, recall, F-measure, and ROC AUC score. The strength of this model lies in its ability to combine diverse decision trees by randomizing cut points and training samples. This unique combination harnesses collective insights, significantly enhancing the model's ability to generalize and make accurate predictions. The ensemble structure of ET played a crucial role in combating overfitting by aggregating predictions from a variety of trees, ensuring stable and reliable performance. Consequently, the ET model was not only proven to be

accurate but also resilient in its predictions, making it a highly promising choice for disease prediction tasks.

When comparing the performance of RF, GB, AdaBoost, DT, KNN, and SVM to the ET model, certain shortcomings became apparent. RF, despite its ensemble nature similar to ET, exhibited a slightly higher susceptibility to overfitting due to its bootstrapping approach. GB's sequential error correction led to overfitting issues, impacting its overall generalization. The iterative misclassification correction of AdaBoost hindered its ability to handle complexities within the dataset, which had a negative impact on its overall performance. DTs tended to overfit, making them less robust compared to ensemble methods. KNN struggled with high-dimensional spaces due to its reliance on local patterns, which affected its generalization. SVM's effectiveness was limited by its dependence on kernel functions and optimal hyperplane separation, particularly when dealing with nonlinear data relationships. These factors combined resulted in the models falling short of achieving the higher performance demonstrated by the ET model. The ET model excels in mitigating overfitting, leveraging diverse insights, and ensuring stable and accurate predictions in disease prediction tasks.

This study introduces a promising diagnostic model designed to identify adults at risk of dementia or heart failure. Future investigations will aim to broaden the dataset, ensuring a more comprehensive representation of the population to improve the accuracy and generalizability of the model. Additionally, collaborative feature engineering techniques will be explored to further enhance the performance of the diagnostic model. The findings emphasize the effectiveness of the ET-based classification model in early detection and differentiation of these conditions. The superior performance of the model not only demonstrates its potential in enhancing diagnostic accuracy but also highlights its role in guiding clinical decision-making. Moving forward, the focus will be on addressing limitations and expanding the capabilities of the model to advance disease classification and proactive early detection using blood test data.

The ET model's superior performance has significant implications for early disease detection. Accurate classification can facilitate timely interventions, resulting in improved patient outcomes. However, it is essential to acknowledge the study's limitations, such as relying on data from a single hospital, which may restrict the model's generalizability to broader populations. Moreover, the use of ICD-10 coding for dementia may not capture

the condition's various subclasses due to insufficient data, thus limiting the model's ability to predict diverse diseases. Future research should expand the dataset by including data from multiple hospitals or different regions to enhance the model's representativeness. Additionally, exploring advanced feature engineering techniques and incorporating additional data sources could optimize the model's performance and enhance its practical application.

CHAPTER 5

CONCLUSIONS

This study proposes an ET classification model for the early detection and categorization of dementia and heart failure. These diseases are commonly correlated among adults, and the model utilizes blood test data. The dataset, sourced from Chiang Rai Prachanukroh Hospital, comprises records from 4,297 individuals, including health assessments and blood test results. The ET algorithm was selected as the classification model due to its capability to effectively handle complex and high-dimensional data, making it a suitable choice for this study. A main contribution of this study is the demonstration that SMOTENC, an oversampling technique designed for both categorical and numerical data, outperforms the widely used CTGAN method. This finding is particularly significant for the classification of heart failure and dementia, where data scarcity can limit the effectiveness of predictive models. This improvement is attributed to SMOTENC's effective methodology for finding the nearest points during oversampling, which enhances the model's performance and leads to more accurate classifications. By applying SMOTENC, the model's performance improved significantly, leading to more accurate classifications. The proposed model underwent evaluation by comparing it against several existing methods, namely KNN, DT, SVM, RF, GB, and AdaBoost. The comparison was based on various evaluation metrics: accuracy, precision, recall, F-measure, and ROC AUC. The results revealed that the ET model outperformed the other models, demonstrating significant improvement in all metrics. This affirms its superiority in accurately classifying dementia and heart failure.

REFERENCES

REFERENCES

[1] Rudnicka, E., Napierała, P., Podfigurna, A., Męczekalski, B., Smolarczyk, R., & Grymowicz, M. (2020). The World Health Organization (WHO) approach to healthy ageing. *Maturitas, 139*, 6-11. https://doi.org/10.1016/j.maturitas.2020.05.018

[2] Jaul, E., & Barron, J. (2017). Age-related diseases and clinical and public health implications for the 85 years old and over population. *Frontiers in Public Health, 5*, 335. https://doi.org/10.3389/fpubh.2017.00335

[3] Cross, P. I. (2023). *Coronary heart disease diagnosis before age 45 may increase dementia risk by 36%*. https://www.medicalnewstoday.com/articles/coronary-heart-disease-early-diagnosis-age-45-increase-dementia-risk#Heart-health-important-when-evaluating-dementia-risk

[4] Wolters, F. J., Segufa, R. A., Darweesh, S. K., Bos, D., Ikram, M. A., Sabayan, B., . . . Sedaghat, S. (2018). Coronary heart disease, heart failure, and the risk of dementia: A systematic review and meta-analysis. *Alzheimer's & Dementia, 14*(11), 1493-1504. https://doi.org/10.1016/j.jalz.2018.01.007

[5] Yap, N. L. X., Kor, Q., Teo, Y. N., Teo, Y. H., Syn, N. L., Evangelista, L. K. M., . . . Sia, C. H. (2022). Prevalence and incidence of cognitive impairment and dementia in heart failure–a systematic review, meta-analysis and meta-regression. *Hellenic Journal of Cardiology, 67*, 48-58. https://doi.org/10.1016/j.hjc.2022.07.005

[6] Arvanitakis, Z., & Bennett, D. A. (2019). What is dementia?. *JAMA, 322*(17), 1728-1728. https://doi.org/10.1001/jama.2019.11653

[7] Yang, H., & Bath, P. A. (2019). The use of data mining methods for the prediction of dementia: Evidence from the English longitudinal study of aging. IEEE *Journal of Biomedical and Health Informatics, 24*(2), 345-353. https://doi.org/10.1109/JBHI.2019.2921418

[8] Arvanitakis, Z., Shah, R. C., & Bennett, D. A. (2019). Diagnosis and management of dementia: Review. *JAMA, 322*(16), 1589. https://doi.org/10.1001/jama.2019.4782

[9] Ryu, S. E., Shin, D. H., & Chung, K. (2020). Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization. *IEEE Access, 8*, 177708-177720. https://doi.org/10.1109/ACCESS.2020.3025553

[10] H Ferreira-Vieira, T., M Guimaraes, I., R Silva, F., & M Ribeiro, F. (2016). Alzheimer's disease: Targeting the cholinergic system. *Current Neuropharmacology, 14*(1), 101-115. https://doi.org/10.2174/1570159X13666150716165726

[11] Harrington, D., Lenahan, C. M., & Beacom, R. (2023). Heart failure management: Updated guidelines. Understand your role in patient-centered care. *American Nurse Journal, 18*(5), 6-12. https://doi.org/10.51256/ANJ052306

[12] Guha, K., & McDonagh, T. (2013). Heart failure epidemiology: European perspective. *Current Cardiology Reviews, 9*(2), 123-127. https://doi.org/10.2174/1573403X11309020005

[13] Berliner, D., Hänselmann, A., & Bauersachs, J. (2020). The treatment of heart failure with reduced ejection fraction. *Deutsches Ärzteblatt International, 117*(21), 376. https://doi.org/10.3238/arztebl.2020.0376

[14] Bader, F., Atallah, B., Brennan, L. F., Rimawi, R. H., & Khalil, M. E. (2017). Heart failure in the elderly: Ten peculiar management considerations. *Heart Failure Reviews, 22*, 219-228. https://doi.org/10.1007/s10741-017-9598-3

[15] Watson, R. D. S. (2000). ABC of heart failure: clinical features and complications. *BMJ, 320*(7229), 236-239. https://doi.org/10.1136/bmj.320.7229.236

[16] Ramani, G. V., Uber, P. A., & Mehra, M. R. (2010). Chronic heart failure: Contemporary diagnosis and management. *Mayo Clinic Proceedings, 85*(2), 180-195. https://doi.org/10.4065/mcp.2009.0494

[17] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., . . . Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access, 10*, 8529-8538. https://doi.org/10.1109/ACCESS.2022.3142097

[18] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., . . . Bolshev, V. (2021). Prediction of chronic kidney disease–a machine learning perspective. *IEEE Access, 9*, 17312-17334. https://doi.org/10.1109/ACCESS.2021.3053763

[19] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[20] Hakim, A., Ng, & Turek. (2013). Heart disease as a risk factor for dementia. *Clinical Epidemiology, 5*(1), 135-145. https://doi.org/10.2147/CLEP.S30621

[21] Winston, C. N., Goetzl, E. J., Akers, J. C., Carter, B. S., Rockenstein, E. M., Galasko, D., . . . Rissman, R. A. (2016). Prediction of conversion from mild cognitive impairment to dementia with neuronally derived blood exosome protein profile. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 3*(1), 63-72. https://doi.org/10.1016/j.dadm.2016.04.001

[22] Alves, T. C. T. F., Rays, J., Fráguas, R., Wajngarten, M., Meneghetti, J. C., Prando, S., . . . Busatto, G. F. (2005). Localized cerebral blood flow reductions in patients with heart failure: A study using 99mTc-HMPAO SPECT. *Journal of Neuroimaging, 15*(2), 150-156. https://doi.org/10.1111/j.1552-6569.2005.tb00300.x

[23] Gruhn, N., Larsen, F. S., Boesgaard, S., Knudsen, G. M., Mortensen, S. A., Thomsen, G., . . . Aldershvile, J. (2001). Cerebral blood flow in patients with chronic heart failure before and after heart transplantation. *Stroke, 32*(11), 2530-2533. https://doi.org/10.1161/hs1101.098360

[24] Hjelm, C., Broström, A., Dahl, A., Johansson, B., Fredrikson, M., & Strömberg, A. (2014). Factors associated with increased risk for dementia in individuals age 80 years or older with congestive heart failure. *Journal of Cardiovascular Nursing, 29*(1), 82-90. https://doi.org/10.1097/JCN.0b013e318275543d

[25] Ahmed, M. R., Zhang, Y., Feng, Z., Lo, B., Inan, O. T., & Liao, H. (2018). Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical Engineering, 12*, 19-33. https://doi.org/10.1590/2446-4740.08117

[26] Zeighami, Y., Fereshtehnejad, S. M., Dadar, M., Collins, D. L., Postuma, R. B., Mišić, B., . . . Dagher, A. (2019). A clinical-anatomical signature of Parkinson's disease identified with partial least squares and magnetic resonance imaging. *Neuroimage, 190*, 69-78. https://doi.org/10.1016/j.neuroimage.2017.12.050

[27] Hah, D. W., Kim, Y. M., & Ahn, J. J. (2019). A study on KOSPI 200 direction forecasting using XGBoost model. *The Korean Data & Information Science Society, 30*(3), 655-669. https://doi.org/10.7465/jkdi.2019.30.3.655

[28] Armananzas, R., Iglesias, M., Morales, D. A., & Alonso-Nanclares, L. (2016). Voxel-based diagnosis of Alzheimer's disease using classifier ensembles. *IEEE Journal of Biomedical and Health Informatics, 21*(3), 778-784. https://doi.org/10.1109/JBHI.2016.2538559

[29] Sharma, S., Gupta, S., Gupta, D., Altameem, A., Saudagar, A. K. J., Poonia, R. C., . . . Nayak, S. R. (2022). HTLML: Hybrid AI based model for detection of Alzheimer's disease. *Diagnostics, 12*(8), 1833. https://doi.org/10.3390/diagnostics12081833

[30] Lombardi, G., Crescioli, G., Cavedo, E., Lucenteforte, E., Casazza, G., Bellatorre, A. G., . . . Filippini, G. (2020). Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD009628.pub2

[31] Buckner, R. L., Snyder, A. Z., Sanders, A. L., Raichle, M. E., & Morris, J. C. (2000). Functional brain imaging of young, nondemented, and demented older adults. *Journal of Cognitive Neuroscience, 12*(Supplement 2), 24-34. https://doi.org/10.1162/089892900564046

[32] Nadel, J., McNally, J. S., DiGiorgio, A., & Grandhi, R. (2021). Emerging utility of applied magnetic resonance imaging in the management of traumatic brain injury. *Medical Sciences, 9*(1), 10. https://doi.org/10.3390/medsci9010010

[33] 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

[34] Armañanzas, R. (2012). *Consensus policies to solve bioinformatic problems: Through bayesian network classifiers and estimation of distribution algorithms*. LAP LAMBERT Academic Publishing.

[35] Kerexeta, J., Larburu, N., Escolar, V., Lozano-Bahamonde, A., Macía, I., Beristain Iraola, A., . . . Graña, M. (2023). Prediction and analysis of heart failure decompensation events based on telemonitored data and artificial intelligence methods. *Journal of Cardiovascular Development and Disease, 10*(2), 48. https://doi.org/10.3390/jcdd10020048

[36] Mavrogiorgou, A., Kiourtis, A., Kleftakis, S., Mavrogiorgos, K., Zafeiropoulos, N., & Kyriazis, D. (2022). A catalogue of machine learning algorithms for healthcare risk predictions. *Sensors, 22*(22), 8615. https://doi.org/10.3390/s22228615

[37] Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., . . . Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access, 7*, 54007-54014. https://doi.org/10.1109/ACCESS.2019.2909969

[38] Angelillo, M. T., Balducci, F., Impedovo, D., Pirlo, G., & Vessio, G. (2019). Attentional pattern classification for automatic dementia detection. *IEEE Access, 7*, 57706-57716. https://doi.org/10.1109/ACCESS.2019.2913685

[39] Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. *IEEE Journal of Biomedical and Health Informatics, 18*(6), 1750-1756. https://doi.org/10.1109/JBHI.2014.2337752

[40] Drvenica, I. T., Stančić, A. Z., Maslovarić, I. S., Trivanović, D. I., & Ilić, V. L. (2022). Extracellular hemoglobin: Modulation of cellular functions and pathophysiological effects. *Biomolecules, 12*(11), 1708. https://doi.org/10.3390/biom12111708

[41] Zhong, X., Na, Y., Yin, S., Yan, C., Gu, J., Zhang, N., . . . Geng, F. (2023). Cell membrane biomimetic nanoparticles with potential in treatment of Alzheimer's disease. *Molecules, 28*(5), 2336. https://doi.org/10.3390/molecules28052336

[42] Li, X., Tan, C., Zhang, W., Zhou, J., Wang, Z., Wang, S., . . . Wei, L. (2015). Correlation between platelet and hemoglobin levels and pathological characteristics and prognosis of early-stage squamous cervical carcinoma. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 21*, 3921. https://doi.org/10.12659/MSM.895016

[43] Santos, G. A. A. D., & Pardi, P. C. (2020). Biomarkers in Alzheimer's disease: Evaluation of platelets, hemoglobin and vitamin B12. *Dementia & Neuropsychologia, 14*, 35-40. https://doi.org/10.1590/1980-57642020dn14-010006

[44] Murdaca, G., Banchero, S., Casciaro, M., Tonacci, A., Billeci, L., Nencioni, A., . . . Gangemi, S. (2022). Potential predictors for cognitive decline in vascular dementia: A machine learning analysis. *Processes, 10*(10), 2088. https://doi.org/10.3390/pr10102088

[45] Rustam, F., Aslam, N., De La Torre Díez, I., Khan, Y. D., Mazón, J. L. V., Rodríguez, C. L., . . . Ashraf, I. (2022, November). White blood cell classification using texture and RGB features of oversampled microscopic images. *Healthcare, 10*(11), 2230. https://doi.org/10.3390/healthcare10112230

[46] Yilmaz, M., Tenekecioglu, E., Arslan, B., Bekler, A., Ozluk, O. A., Karaagac, K., . . . Akgumus, A. (2015). White blood cell subtypes and neutrophil–lymphocyte ratio in prediction of coronary thrombus formation in non-ST-segment elevated acute coronary syndrome. *Clinical and Applied Thrombosis/Hemostasis, 21*(5), 446-452. https://doi.org/10.1177/1076029613507337

[47] Cisternas, P., Lindsay, C. B., Salazar, P., Silva-Alvarez, C., Retamales, R. M., Serrano, F. G., . . . Inestrosa, N. C. (2015). The increased potassium intake improves cognitive performance and attenuates histopathological markers in a model of Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1852*(12), 2630-2644. https://doi.org/10.1016/j.bbadis.2015.09.009

[48] Wallig, M. A., Bolon, B., Haschek, W. M., & Rousseaux, C. G. (Eds.). (2017). *Fundamentals of toxicologic pathology*. Academic press.

[49] Deeks, A., Lombard, C., Michelmore, J., & Teede, H. (2009). The effects of gender and age on health related behaviors. *BMC Public Health, 9*(1), 1-8. https://doi.org/10.1186/1471-2458-9-213

[50] Schäfer, I., Hansen, H., Schön, G., Höfels, S., Altiner, A., Dahlhaus, A., . . . Wiese, B. (2012). The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. First results from the multicare cohort study. *BMC Health Services Research, 12*(1), 1-15. https://doi.org/10.1186/1472-6963-12-89

[51] Shen, F. X., Wolf, S. M., Bhavnani, S., Deoni, S., Elison, J. T., Fair, D., . . . Vaughan, J. T. (2021). Emerging ethical issues raised by highly portable MRI research in remote and resource-limited international settings. *Neuroimage, 238*, 118210. https://doi.org/10.1016/j.neuroimage.2021.118210

[52] Dairi, A., Harrou, F., & Sun, Y. (2021). Deep generative learning-based 1-SVM detectors for unsupervised COVID-19 infection detection using blood tests. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-11. https://doi.org/10.1109/TIM.2021.3130675

[53] Gao, X., & Li, G. (2020). A KNN model based on Manhattan distance to identify the SNARE proteins. *IEEE Access, 8*, 112922-112931. https://doi.org/10.1109/ACCESS.2020.3003086

[54] Huynh-Cam, T. T., Chen, L. S., & Le, H. (2021). Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. *Algorithms, 14*(11), 318. https://doi.org/10.3390/a14110318

[55] Battista, K., Patte, K. A., Diao, L., Dubin, J. A., & Leatherdale, S. T. (2022). Using decision trees to examine environmental and behavioural factors associated with youth anxiety, depression, and flourishing. *International Journal of Environmental Research and Public Health, 19*(17), 10873. https://doi.org/10.3390/ijerph191710873

[56] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access, 8*, 107562-107582. https://doi.org/10.1109/ACCESS.2020.3001149

[57] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. Vieira (eds.), *Machine learning: Methods and applications to brain disorders* (pp. 101-121). Academic Press.

[58] Guo, X., & Hao, P. (2021). Using a random forest model to predict the location of potential damage on asphalt pavement. *Applied Sciences, 11*(21), 10396. https://doi.org/10.3390/app112110396

[59] Purwanto, A. D., Wikantika, K., Deliar, A., & Darmawan, S. (2022). Decision tree and random forest classification algorithms for mangrove forest mapping in Sembilang National Park, Indonesia. *Remote Sensing, 15*(1), 16. https://doi.org/10.3390/rs15010016

[60] Wang, D., Hwang, J., Lee, J., Kim, M., & Lee, I. (2023). Temperature-based state-of-charge estimation using neural networks, gradient boosting machine and a jetson nano device for batteries. *Energies, 16*(6), 2639. https://doi.org/10.3390/en16062639

[61] Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An efficient AdaBoost algorithm with the multiple thresholds classification. *Applied Sciences, 12*(12), 5872. https://doi.org/10.3390/app12125872

[62] Olatunji, S. O., Alsheikh, N., Alnajrani, L., Alanazy, A., Almusairii, M., Alshammasi, S., . . . Alhiyafi, J. (2023). Comprehensible machine-learning-based models for the pre-emptive diagnosis of multiple sclerosis using clinical data: A retrospective study in the eastern province of Saudi Arabia. *International Journal of Environmental Research and Public Health, 20*(5), 4261. https://doi.org/10.3390/ijerph20054261

[63] Aldossary, Y., Ebrahim, M., & Hewahi, N. (2022). A comparative study of heart disease prediction using tree-based ensemble classification techniques. In *2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 353-357). IEEE.

[64] Utari, D. T. (2023). Integration of SVM And Smote-NC for Classification of Heart Failure Patients. *Barekeng: Jurnal Ilmu Matematika Dan Terapan, 17*(4), 2263-2272. https://doi.org/10.30598/barekengvol17iss4pp2263-2272

[65] Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation, 4*(1), 18. https://doi.org/10.3390/asi4010018

[66] Majeed, A., & Hwang, S. O. (2023). CTGAN-MOS: Conditional generative adversarial network based minority-class-augmented oversampling scheme for imbalanced problems. *IEEE Access, 11*, 85878-85899. https://doi.org/10.1109/ACCESS.2023.3303509

[67] Eom, G., & Byeon, H. (2023). Searching for optimal oversampling to process imbalanced data: Generative adversarial networks and synthetic minority over-sampling technique. *Mathematics, 11*(16), 3605. https://doi.org/10.3390/math11163605

[68] García-Vicente, C., Chushig-Muzo, D., Mora-Jiménez, I., Fabelo, H., Gram, I. T., Løchen, M. L., . . . Soguero-Ruiz, C. (2023). Evaluation of synthetic categorical data generation techniques for predicting cardiovascular diseases and post-hoc interpretability of the risk factors. *Applied Sciences, 13*(7), 4119. https://doi.org/10.3390/app13074119

[69] Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information, 11*(6), 332. https://doi.org/10.3390/info11060332

[70] Madni, H. A., Umer, M., Abuzinadah, N., Hu, Y. C., Saidani, O., Alsubai, S., . . . Ashraf, I. (2023). Improving sentiment prediction of textual Tweets using feature fusion and deep machine ensemble model. *Electronics, 12*(6), 1302. https://doi.org/10.3390/electronics12061302

[71] Beam, C. R., Kaneshiro, C., Jang, J. Y., Reynolds, C. A., Pedersen, N. L., & Gatz, M. (2018). Differences between women and men in incidence rates of dementia and Alzheimer's disease. *Journal of Alzheimer's Disease, 64*(4), 1077-1083. https://doi.org/10.3233/JAD-180141

[72] Alzheimer's Association. (2014). 2014 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 10*(2), e47-e92. https://doi.org/10.1016/j.jalz.2014.02.001

[73] Chou, A. F., Wong, L., Weisman, C. S., Chan, S., Bierman, A. S., Correa-de-Araujo, R., . . . Scholle, S. H. (2007). Gender disparities in cardiovascular disease care among commercial and Medicare managed care plans. *Women's Health Issues, 17*(3), 139-149. https://doi.org/10.1016/j.whi.2007.03.004

[74] Bozkurt, B., & Khalaf, S. (2017). Heart failure in women. *Methodist DeBakey Cardiovascular Journal, 13*(4), 216. https://doi.org/10.14797/mdcj-13-4-216

[75] Shah, R., & Agarwal, A. K. (2013). Anemia associated with chronic heart failure: current concepts. *Clinical Interventions in Aging, 8*, 111-122. https://doi.org/10.2147/CIA.S27105

[76] Wolters, F. J., Zonneveld, H. I., Licher, S., Cremers, L. G., Ikram, M. K., Koudstaal, P. J., . . . Ikram, M. A. (2019). Hemoglobin and anemia in relation to dementia risk and accompanying changes on brain MRI. *Neurology, 93*(9), e917-e926. https://doi.org/10.1212/WNL.0000000000008003

[77] Zaidi A. (2023). Predicting wildfires in Algerian forests using machine learning models. *Heliyon, 9*(7), e18064. https://doi.org/10.1016/j.heliyon.2023.e18064

[78] Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules, 26*(4), 1111. https://doi.org/10.3390/molecules26041111

[79] Chollet, F. (2017). *Deep learning with Python*. Manning Publications.

APPENDICES

APPENDIX A

DATA DISTRIBUTION AND EXPLORATION

A1. Disease Prevalence

According to the dataset, the number of patients with both diseases is identical in 4,297 patient records with ten features. The ICD-10 codes in the PDX feature were derived, and dementia was labeled with "Dementia" and heart failure with "Heart Failure". This feature was used to identify the information for each related feature, which comprises the data in Figure 3.2. The disease ratio is shown in Figure A1.

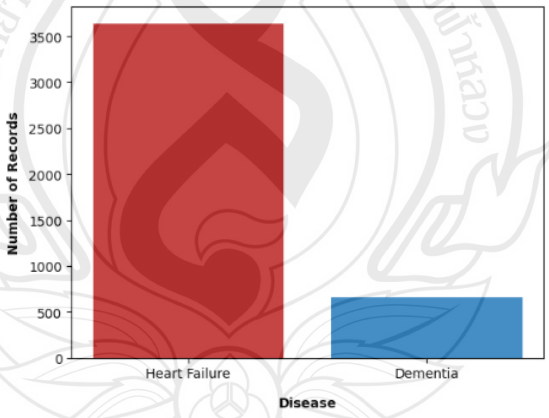

Figure A1 Disease Prevalence Ratios

A2. A Comparison of Gender

For gender, the ratio of male to female prevalence of each disease is shown in Figure A2. Females are 59.72% more likely than males to have dementia, while they are 56.16% more likely to experience heart failure.

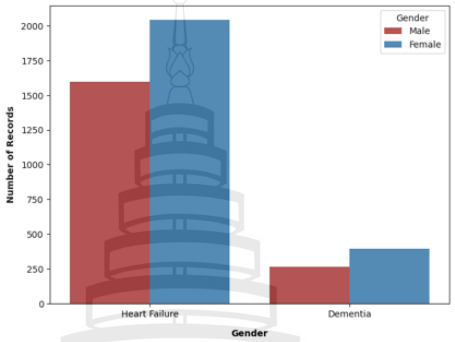

Figure A2 Comparison of Gender Ratios

A3. Age Distribution

As shown in Figure A3, this study examines the prevalence of dementia or heart failure in older adults. Data were collected from patients over 60, with data volume similar to the age of 80.

Figure A3 Age Distribution

A4. Patients with Abnormal Creatinine Levels

As shown in Figure A4, creatinine levels should be between 0.6 and 1.3 milligrams per deciliter (mg/dL) for adults. Abnormal creatinine levels were found in only 152 of 599 dementia patients, accounting for 25.37%. For heart failure, there was nearly double the number of patients with abnormal creatinine values, 60.76%, or 2,170 out of 3,571 patients.

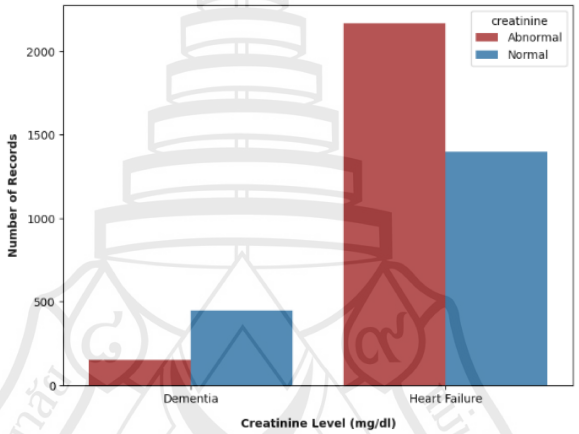

Figure A4 Abnormal Creatinine Levels in Patients

A5. Patients with Abnormal Blood Urea Nitrogen Levels

As shown in Figure A5, blood urea nitrogen levels are measured in milligrams per deciliter (mg/dL), and levels higher than 25 mg/dL are considered abnormal. There were 359 dementia patients with normal blood urea nitrogen levels, with only 76 for abnormal patients, or 17.47% of patients with abnormal levels. However, more than 54.12% of heart failure patients had abnormal values.

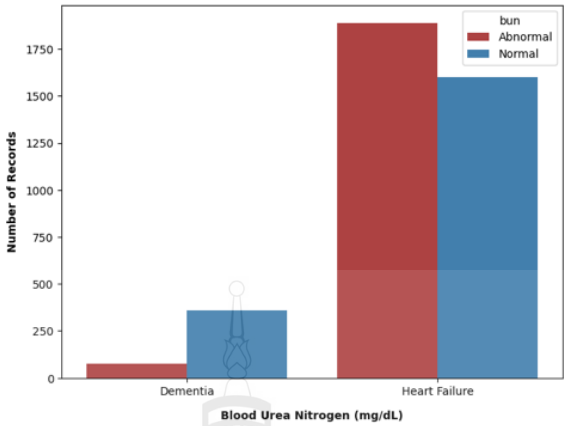

Figure A5 Abnormal Blood Urea Nitrogen Levels in Patients

A6. Patients with Abnormal Hemoglobin Levels

As shown in Figure A6, hemoglobin (Hb) is a protein found in red blood cells re-sponsible for oxygen delivery to tissues. The criterion for the normal range of hemoglobin concentration is generally 16.0 grams per deciliter (g/dL). Hemoglobin levels were abnormal in 53.04% of dementia patients and 64.35% of heart failure patients. This proportion may indicate a significant association with the diseases.

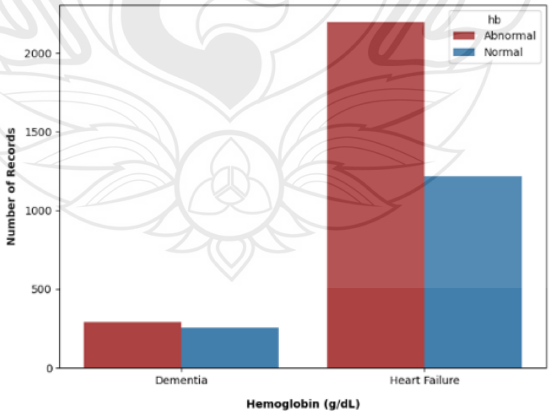

Figure A6 Abnormal Hemoglobin Levels in Patients

A7. Patients with Abnormal Potassium Levels

As shown in Figure A7, a typical adult potassium level ranges between 3.5 and 5.3 millimoles per liter (mmol/L). Only 19.47% of dementia patients had abnormal levels, suggesting that they may have a lesser impact than in heart failure, where 31.23% of patients had abnormal values.

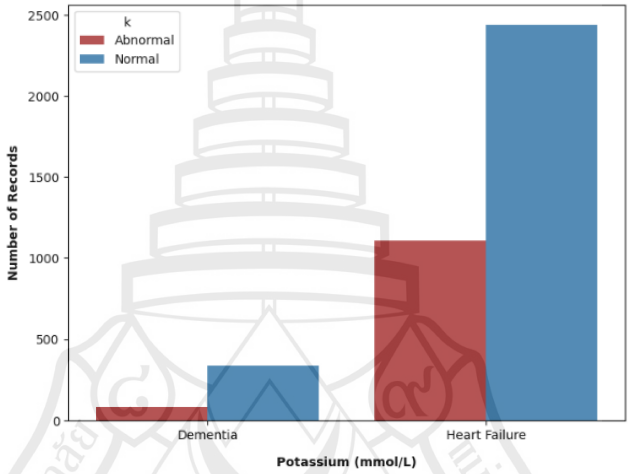

Figure A7 Abnormal Potassium Levels in Patients

A8. Patients with Abnormal White Blood Cells Levels

As shown in Figure A8, white blood cells in the blood are considered normal if the level ranges between 4,000 to 11,000 per microliter. According to the dataset, this feature impacted 13.67% of dementia patients and 32.85% of heart failure patients.

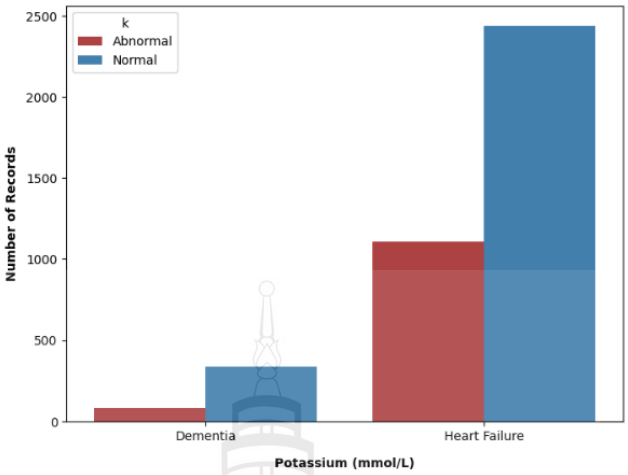

Figure A8 Abnormal White Blood Cells Levels in Patients

A9. Patients with Abnormal Neutrophil Levels Across Diseases

As shown in Figure A9, a normal, healthy neutrophil count should be between 2,500 and 7,000 neutrophils per microliter of blood. Abnormal readings were found in 56.24% of heart failure patients and 17.37% of dementia patients.

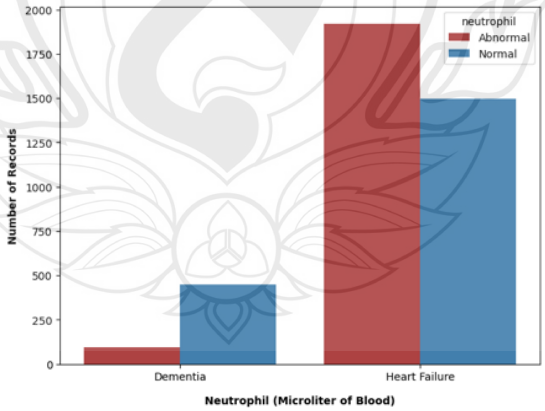

Figure A9 Abnormal Neutrophil Levels in Patients

A10. Patients with Abnormal Platelet Levels Across Diseases

As shown in Figure A10, a normal platelet count ranges from 150,000 to 400,000 platelets per microliter of blood. Platelet abnormalities were found in 9.42% of dementia patients and 22.61% of heart failure patients.

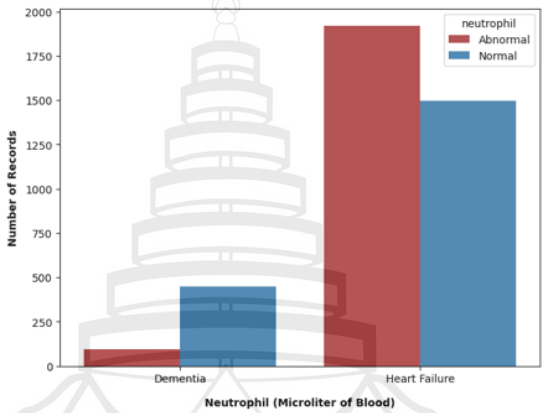

Figure A10 Abnormal Platelet Levels in Patients

A11. Patients with Abnormal Lymphocyte Levels

As shown in Figure A11, the usual range of human lymphocytes is 1,000 to 4,800 per microliter of blood. Of the 541 patients with dementia disease, only 17 patients had white blood cell abnormalities, and 941 patients with heart failure had the same abnormalities out of 3,416 patients.

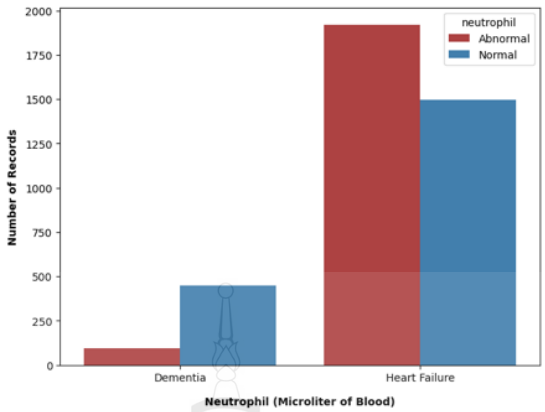

Figure A11 Abnormal Lymphocyte Levels in Patients

APPENDIX B

COMPREHENSIVE PERFORMANCE OF PCA RESULTS

Table B1 Performance Metrics for PCA with 3 Principal Components

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	73.62%	92.06%	76.46%	83.54%	65.07%
GB	67.98%	94.90%	67.02%	78.56%	70.88%
Ada	66.27%	94.76%	65.07%	77.16%	69.90%
RF	73.75%	91.92%	76.76%	83.66%	64.70%
DT	70.34%	90.02%	74.36%	81.44%	58.23%
KNN	70.21%	92.97%	71.36%	80.75%	66.73%
SVM	61.42%	96.28%	58.17%	72.52%	71.19%

Table B2 Performance Metrics for PCA with 5 Principal Components

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	80.18%	92.02%	84.71%	88.21%	66.56%
GB	74.80%	94.73%	75.41%	83.97%	72.97%
Ada	73.88%	94.66%	74.36%	83.29%	72.44%
RF	79.53%	92.65%	83.21%	87.68%	68.45%
DT	75.20%	91.07%	79.46%	84.87%	62.36%
KNN	75.07%	93.60%	76.76%	84.35%	69.96%
SVM	66.93%	95.40%	65.37%	77.58%	71.63%

Table B3 Performance Metrics for PCA with 7 Principal Components

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	84.78%	93.39%	88.91%	91.09%	72.35%
GB	79.13%	95.20%	80.21%	87.06%	75.89%
Ada	75.72%	94.96%	76.31%	84.62%	73.95%
RF	82.94%	93.24%	86.81%	89.91%	71.30%
DT	79.53%	93.83%	82.01%	87.52%	72.06%
KNN	76.90%	94.88%	77.81%	85.50%	74.17%
SVM	71.13%	96.08%	69.87%	80.90%	74.93%

Table B4 Performance Metrics for PCA with 9 Principal Components

Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ET	87.14%	93.17%	92.05%	92.61%	72.34%
GB	80.31%	95.27%	81.56%	87.88%	76.57%
Ada	76.12%	95.16%	76.61%	84.88%	74.62%
RF	85.56%	93.31%	89.96%	91.60%	72.35%
DT	79.92%	92.13%	84.26%	88.02%	66.87%
KNN	75.98%	94.98%	76.61%	84.81%	74.10%
SVM	72.57%	96.36%	71.36%	82.00%	76.21%

APPENDIX C

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: 148/2023 **Protocol No: EC 23130-13**

Title: classification of dementia and heart failure in adult people

Principal investigator: Mr.pornthep Phanbua

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: August 17, 2023

Sangkae

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

For research protocol exempted by the Mae Fah Luang University Ethics Committee on Human Research (MFU EC), the investigators must comply with the followings:

- No need to submit a progress report.
- When there are changes of the protocol, the investigator must send an amendment report (AP 06/2022) to the MFU EC.
- When the research finishes, the investigator must send a final report (AP 09/2022).

Please go to https://ec.mfu.ac.th to download MFU EC forms for reporting.

I, as an investigator, agree to comply with the above obligation.

Pornthep

(Mr.pornthep Phanbua)
Date 14/09/2023

CURRICULUM VITAE

CURRICULUM VITAE

NAME Pornthep Phanbua

EDUCATIONAL BACKGROUND

2021 Bachelor of Science
Software Engineering
Mae Fah Luang University