# A Lifestyle Related Disease Prediction Framework Based on Missing Value Imputation and Stacking Ensemble Method

Lijuan Ren

N° d'ordre NNT : 2023LYO20010

# THÈSE de DOCTORAT DE L'UNIVERSITÉ LUMIÈRE LYON 2

**École Doctorale : ED 512 Informatique et Mathématiques**

Discipline : Informatique

Soutenue publiquement le 16 mars 2023, par :

Lijuan REN

# A Lifestyle Related Disease Prediction Framework Based on Missing Value Imputation and Stacking Ensemble Method.

Devant le jury composé de :

Hervé PINGAUD, Professeur des Universités, Institut National Universitaire Champollion, Président
Lina SOUALMIA, Professeure des Universités, Université Rouen Normandie, Rapporteure
Keshav DAHAL, Professeur des Universités, University of the West of Scotland, Rapporteur
Haiqing ZHANG, Professeure, Chengdu University of Information Techno, Examinatrice
Tewfik ZIADI, Maître de conférences HDR, Sorbonne Université, Examinateur
Tao WANG, Maître de conférences HDR, Université Jean Monnet Saint-Étienne, Examinateur
Aïcha SEKHARI, Maîtresse de conférences, Université Lumière Lyon 2, Examinatrice
Abdelaziz BOURAS, Professeur des Universités, Université Lumière Lyon 2, Directeur de thèse

# Contrat de diffusion

# THESE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

UNIVERSITÉ LUMIÈRE LYON 2

Ecole Doctorale Informatique et Mathématiques de Lyon / ED 512 Laboratoire Décision et Information pour les Systèmes de Production / EA 4570

---

# A Risk Prediction Framework for Lifestyle Related Disease Based on Missing Values Imputation and Stacking Ensemble Method

---

Soutenue par


## Lijuan REN


Pour obtenir le grade de : DOCTEUR EN INFORMATIQUE

Sous la direction de : Prof. Abdelaziz BOURAS, Dr. Aicha SEKHARI, Dr. Tao WANG

Devant un Jury composé de :

| | |
|---|---|
| Mme. Lina SOUALMIA, Professeur, Université de Rouen, | Rapporteur |
| M. Keshav DAHAL, Professeur, University of the West of Scotland, | Rapporteur |
| M. Tewfik ZIADI, Maître de Conférences-HDR, Sorbonne Université, | Examinateur |
| M. Hervé PINGAUD, Professeur, Institut National Universitaire Champollion, | Examinateur |
| Mme. Haiqing ZHANG, Professeur, Chengdu University of Information Technology, China, | Examinateur |
| M. Abdelaziz BOURAS, Professeur des Universités, Université Lumière Lyon 2, | Directeur de thèse |
| Mme. Aicha SEKHARI, Maître de Conférences, Université Lumière Lyon2, | Co-encadrant de thèse |
| M. Tao WANG, Maître de Conférences-HDR, Université Jean Monnet de Saint-Etienne | Co-encadrant de thèse |

# ABSTRACT

Industrialized countries have come to the conclusion that numerous chronic non-communicable diseases are caused by lifestyle-related factors after completing numerous epidemiological studies on these conditions, and can be called lifestyle-related diseases (LRDs). Obesity, high blood pressure, coronary heart disease, and other cardiovascular diseases, stroke and other cerebrovascular diseases, diabetes, and several malignant tumors are among the diseases that are included. All of these conditions pose a major threat to people's lives and health and are challenging to treat with current medical technology.

In this context, the prevention of lifestyle-related diseases is extremely important. Disease prediction facilitates early detection to improve the chances of positive health outcomes. Therefore, this study aims to propose a lifestyle-related disease prediction framework based on missing value imputation and stacking ensemble method. Specifically, the application of information technology in the medical field is resulting in a large amount of medical data. However, due to early withdrawal and refusal of participants, there are a lot of missing values in medical data. We proposed an imputation method based on SMOTE-NC oversampling technology and the ALWRF method for imbalanced and mixed-type data, called SncALWRFI. Meanwhile, Bayesian optimization and cross-validation are employed to search optimal parameters. In the experiment for missing value imputation, the SncALWRFI shows the best imputation accuracy, and it performs high imputation effectiveness in public datasets with characteristics of data imbalance and mixed type.

Since prediction performance can be easily impacted by the presence of noise, we have to look for a good strategy to improve this situation. Noise may come from real patients and cannot be removed directly. Meanwhile, ensemble approaches are a great way to lower variation, bias, and noise. Therefore, in order to increase the

prediction performance of lifestyle-related diseases, we employ the stacking ensemble technology in our study. Specifically, in order to maximize the diversity and the accuracy of ensemble models simultaneously, we proposed a Multi-objective Iterative Model Selection (MoItMS) algorithm. Data were obtained from the National Health and Nutrition Examination Survey from 2007 to 2018. Our study utilized an imbalanced data set of 11,341 with (67.16%) non-hypertensive patients, and (32.84%) hypertensive patients. The results indicate a sensitivity of 51.41%, a specificity of 70.48%, an accuracy of 76.62%, and a measured AUC (Area under the ROC Curve) of 0.84, which outperformed 12 individual and ensemble models. The proposed ensemble model can be implemented in applications to assist population health management programs in identifying patients with a high risk of developing hypertension.

The missing value module, feature selection module, and disease prediction module are the three main elements of the architecture we propose for LRDs prediction. In view of the large number of missing values in the data set related to lifestyle-related diseases, the missing value module uses a combination of deletion and imputation to deal with missing values. Since different lifestyle-related diseases have different relevant features, the feature selection module uses machine learning-based feature selection to find key features for lifestyle-related diseases. Finally, we use a scenario from a Chinese hospital to apply the suggested prediction framework. According to the experimental findings, the proposed prediction framework can also enhance LRD's prevention performance.

**Keywords:** Lifestyle-related diseases, Prediction, Machine Learning, Missing values, Stacking ensemble.

# RÉSUMÉ

Les pays industrialisés sont arrivés à la conclusion que de nombreuses maladies chroniques non transmissibles sont causées par des facteurs liés au mode de vie après avoir réalisé de nombreuses études épidémiologiques sur ces conditions, et peuvent être appelées maladies liées au mode de vie (MRD). L'obésité, l'hypertension artérielle, les maladies coronariennes et autres maladies cardiovasculaires, les accidents vasculaires cérébraux et autres maladies cérébrovasculaires, le diabète et plusieurs tumeurs malignes font partie de ces maladies. Toutes ces conditions constituent une menace majeure pour la vie et la santé des personnes et sont difficiles à traiter avec la technologie médicale actuelle.

Dans ce contexte, la prévention des maladies liées au mode de vie est extrêmement importante. La prévention des maladies facilite la détection précoce pour améliorer les chances de résultats positifs pour la santé. Par conséquent, cette étude vise à proposer un cadre de prédiction des maladies liées au mode de vie basé surl'imputation des valeurs manquantes et l'ensemble la méthode ensembliste. Plus précisément, l'application des technologies de l'information dans le domaine médical produit une grande quantité de données médicales. Cependant, à cause de certaines situations de la collecte de données, comme le retrait précoce et le refus des participants, il y a beaucoup de valeurs manquantes dans les données médicales. Nous avons proposé une méthode d'imputation basée sur la technologie de suréchantillonnage SMOTE-NC et la méthode ALWRF pour les données déséquilibrées et de type mixte, appelée SncALWRFI. Pendant ce temps, l'optimisation bayésienne et la validation croisée sont utilisées pour rechercher les paramètres optimaux. Dans l'imputation des valeurs manquantes, le SncALWRFI présente une meilleure précision d'imputation et réalise

une efficacité d'imputation élevée pour l'ensemble des bases de données publiques avec des caractéristiques de déséquilibre et de type mixe.

Étant donné que les performances de prédiction peuvent être facilement impactées par la présence de bruit dans les données, nous devons rechercher une bonne stratégie pour améliorer cette situation. Le bruit peut provenir de vrais patients et il ne peut être supprimé directement.Les approches d'ensemble sont un excellent moyen de réduire la variation, le biais et le bruit. Par conséquent, afin d'augmenter les performances de prédiction des maladies liées au mode de vie, nous utilisons la technologie d'approche ensembliste dans notre étude pour confronter au bruit des données. Plus précisément, afin de maximiser simultanément la diversité et la précision des modèles d'ensemble, nous avons proposé un algorithme multi-objectif de sélection itérative de modèles (MoItMS). Les données ont été obtenues à partir de l'enquête nationale sur la santé et la nutrition de 2007 à 2018. Notre étude a utilisé un ensemble de données déséquilibrées de 11 341 personnes avec (67,16%) personnes non hypertendues et (32,84%) patients hypertendus. Les résultats indiquent une sensibilité de 51,41 %, une spécificité de 70,48 %, une précision de 76,62 % et une AUC mesurée à 0,84, ce qui a surpassé 12 modèles individuels et d'ensemble. Ce modèle peut être mis en œuvre dans des applications pour aider les programmes de santé publique à identifier les patients présentant un risque élevé de développer une hypertension.

Le module de l'imputation de valeur manquante, le module de sélection des caractéristiques et le module de prédiction des maladies sont les trois principaux éléments de l'architecture que nous proposons pour la prédiction des LRD. Pour un grand nombre de valeurs manquantes, la méthode combinant la suppression et l'imputation est sélectionnée comme principale stratégie de traitement des valeurs manquantes. Étant donné que différentes maladies liées au mode de vie ont des caractéristiques différentes, le module de sélection de caractéristiques utilise une méthode basée sur l'apprentissage automatique pour trouver des caractéristiques clés. Enfin,

nous utilisons un scénario chinois pour expérimenter le cadre de prédiction suggéré. Selon les résultats expérimentaux, le cadre de prédiction proposé peut également améliorer les performances d'évaluation des risques de LRD.

**Mots clés:** Maladies liées au mode de vie, Prédiction, Apprentissage automatique, Valeurs manquantes, Ensemble d'empilement.

# Acknowledgements

The study tour in France is a dream-seeking journey for me. During my master's degree, I lived the worst period of my life, and the huge graduation pressure always weighed on me. I seemed to be floating in an endless sea, but this did not extinguish my determination to continue chasing my dream. When I had the opportunity to apply to study in France, I submitted the application without hesitation. It's a bit marvelous to say that my love for France originally came from a picture of a sea of lavender flowers in Provence, France that I saw when I was very young. I am grateful for the sheer desire that France brought to me and made me believe that I could make my dreams come true in this place.

If France was the fertile ground for my studies, then my director Abdelaziz BOURAS provided wings for my doctoral studies. I remember that in the first interview, the network was very poor, I was very nervous, and my performance was not satisfactory. But he always said to me, "it's fine", "it's good", "it's ok". I know he was encouraging me and guiding me into a better state. Later, his encouragement has been with me from the beginning to the end and has never stopped, helping me from the hesitation at the beginning to the firmness at the end. I am really grateful for his guidance and help during my doctoral period. I can always improve when I communicate with him. At the same time, he also taught me many skills besides scientific research, including the way of thinking about problems, the ability to logical analysis, and a healthy lifestyle.

Moreover, I would like to thank my three co-supervisors, Aïcha SEKHARI SEKLOULI, Tao WANG, and Haiqiang ZHANG. First of all, I would like to thank Prof. SEKHARI, for her in-depth sharing of French culture and language learning which enables me to enjoy fully In a foreign environment. At the same time, she

can always provide me with novel research ideas and constantly promote the progress of my research. I also want to thank Prof. WANG for their endless help during these years. He has a wealth of professional knowledge and always provides me with great ideas when I encounter complex research problems. In addition, I am also very grateful to Prof. ZHANG, she is the enlightener of my doctoral study, and she not only provided me with the professional guidance in my study but also shared a lot of work and life experience with me.

I would like to thank Bilgesu BAYIR and Boubou Thiam NIANG, who are not only my colleagues but also my friends. We entered the lab at the same time. They provided a lot of help when I first arrived in France and made me adapt to the study and life in France quickly. The friendships grow stronger with the days and it would be cherished for a lifetime. Thanks to all my friends in the laboratory, Jiao Zhao, Qing Li, Badreddine TANANE, and Mengji Yang, who have brought me enormous joy. I also thank the professors in the laboratory, Vincent CHEUTET, Yacine OUZROUT, Néjib MOALLA, for giving me a lot of help and support. I also want to thank Huiru Ren, for cooking me delicious Chinese food and traveling with me. Thanks to Jianan Xu for their help when I first come to Lyon. Thanks to Yuxin Zong, who helped me a lot during these years. I am also very grateful to He Feng for his help in professional learning.

I want to thank my parents and my family for their unconditional love and support. Of course, I also want to thank myself for giving myself more opportunities to see the beautiful world. At last, I would like to thank the China Scholarship Council (CSC) for providing me with a scholarship to help me study in France smoothly.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

$MSIE_{num}$ Mean of Squared Imputation Errors for numerical values 57

$wNNSel_{mix}$ Weighted Nearest Neighbor Imputation using Selected Variables 66

$PFC_{cat}$ Proportion of Falsely Imputed Categories 57

**ABDT** Adaptive Boosting Decision Tree 70

**AdaBoost** Adaptive Boosting 11

**ALWRF** Adaptive Laplacian Weight Random Forest 29

**AUC** Area under the ROC(Receiver operating characteristic) curve 3

**BMI** Body Mass Index 59

**CART** Classification and Regression Tree 40

**DT** Decision Tree 9, 27

**Gaussian NB** Gaussian Naive Bayes Network 110, 124

**GBDT** Gradient Boosting Decision Tree 9, 70

**kNN** k-Nearest Neighbors 9, 27

**kNNI** k-Nearest Neighbors Imputation 66

**LightGBM** Light Gradient Boosting Machine 12, 124

**LR** Linear Regression 9, 72

**LRDs** Lifestyle-related diseases 3

**LRM** Logistic Regression Model 3, 9, 98

**MLP** Multi-Layer Perceptron 10, 70

**MoItMS** Multi-objective Iterative Model Selection 114

**MSE** Mean Squared Error 55

**NB** Naive Bayes Network 9, 72

**NHANES** National Health and Nutrition Examination Survey 60

**RF** Random Forest 9, 27, 124

**RFI** Random Forest Imputation or MissForest 66

**SHAP** SHapley Additive exPlanations 132, 133

**SMOTE** Synthetic Minority Over-sampling TEchnique 49

**SMOTE-NC** Synthetic Minority Oversampling Technique for Nominal and Continuous 29

**SncALWRFI** SMOTE-NC and ALWRF Imputation 51

**SVM** Support Vector Machine 9, 72

**XGBoost** Extreme Gradient Boosting 10, 124

# Chapter 1. Introduction

## 1.1 Background

Traditional medical services used to be kept and recorded on paper [1, 2], which is difficult to serve people effectively and easily as society has developed [3]. The medical industry has seen revolutionary changes as a result of the digitization of medical information [4]. Through the systematization, standardization, and intelligence of big data, the digital medical system effectively integrates various patient information data, offers intelligent services for patients, and intelligent management based on electronic files for hospitals [5]. The construction of a digital hospital management system is essential in order to improve the operational efficiency of modern hospitals. The database of the hospital information system includes a variety of medical data, including administrative data, laboratory data, treatment data, and prescription data [6]. The amount of data keeps growing over time, and the gathered knowledge about medical practices can serve as a guide for the conduct of the clinical medical staff as well as a wealth of useful information for hospital administrators [7]. Additionally, examining and mining this beneficial data can yield important references for making medical decisions [8].

However, the exponential rise of medical data as a result of the quick development of medical information technology has made its hidden value an urgently

needed treasure. Especially, with the improvement of people's living standards and health awareness, more and more health check data are collected. For example, the "China Health Statistical Yearbook" showed that 444 million health examinations were performed in China in 2019 compared to 406 million in 2017. These health examinations produce enormous amounts of medical data with hidden value. In order to offer people intelligent and individualized medical services, it is urgently necessary to mine the valuable information concealed in massive amounts of medical data [9]. In particular, the use of information-based methods to screen data allows administrators and healthcare professionals to thoroughly research patient medical histories and deliver more effective care [10]. Accurate and individualized health care services can be provided by utilizing big data analysis techniques in the medical and health fields, as well as data mining and analysis technology to examine medical data [6]. In this context, it is crucial to employ big data and artificial intelligence to discover valuable information hidden in massive data sets held in medical information systems and to equip local hospitals with smart medical systems to boost the effectiveness of healthcare services.

### 1.2 Research Significance

As we know, there is a global lack of medical resources, including general practitioners and medical supplies. For example, only 800 doctors were practicing medicine in the French department of Seine-et-Marne as of December 31, 2020, or less than 6 doctors for every 10,000 people [11]. Therefore, more and more researchers use information technology to assist doctors in their work to improve

service efficiency. For example, Mohamed Elhoseny et al.[12] proposed a classification system of chronic kidney disease to help doctors distinguish different groups and achieved a prediction accuracy of 95%. Although it is difficult for these methods to predict all cases perfectly, they can be used as additional tools to provide information to doctors. On the other hand, some studies focus on preventing or delaying the progression of the disease. For example, Shuqiong Huang et al. proposed an artificial neural network method to use risk factors to evaluate the risk of hypertension. Their model achieves 90% Area under the ROC(Receiver operating characteristic) curve (AUC) performance better than Logistic Regression Model (LRM) in assessing HTN risk. Risk evaluation methods have obvious advantages in disease prevention. They predict people's risks based on risk factors before the disease occurs, and assist doctors in providing early intervention, which can reduce medical expenses and people's suffering from diseases.

However, disease prevention approaches have some limitations, and they are more suitable for diseases where risk factors are readily available and disease progression is improvable. Lifestyle-related diseases (LRDs) have natural advantages to building disease risk prediction models. LRDs refer to diseases whose psychophysiology is significantly affected by lifestyle factors, and changes in these etiological factors can significantly improve disease prevention and treatment [13, 14]. From the definition of LRDs, they are extremely related to people's lifestyles or behaviors, their risk factors are easily obtained, and many studies [15, 16] have shown that LRDs can be improved by healthy lifestyles.

On the other hand, as countries become more industrialized and wealthier, the prevalence of LRDs increases due to changes in people's behavior. Generally, most chronic diseases, including cardiovascular disease, metabolic syndrome, obesity, type 2 diabetes, and some cancers, are lifestyle-related diseases and closely related to people's lifestyles [14]. Studies have found that lifestyle-related diseases are the absolute and relative most common diseases in the world today, and the death toll exceeds that of AIDS, malaria, and tuberculosis combined [17]. Cardiovascular disease, obesity, type 2 diabetes, hypertension, and some particular malignancies have all grown to be significant problems in the twenty-first century. In the Republic of Ireland, 61% of adults are overweight or obese, and over 40% of adults report having at least one lifestyle-related disease, the most prevalent of which is high blood pressure and high cholesterol [18]. Additionally, 17.8 million individuals globally passed away from cardiovascular disease (CVD) in 2017, according to the Global Burden of Disease report published in 2018 and the estimated overall number of tumor-related fatalities (mostly cancer) is 9.56 million [19]. The WHO predicts that by 2030, there will be 366 million individuals worldwide who have diabetes, up from the present estimate of 175 million [20]. Despite the availability of a wide range of medicines, the frequency of lifestyle illnesses is not controlled due to the safety concerns connected with these medicines[21]. To sum up, there is a crisis in the global healthcare system as a result of the prevalence of these lifestyle-related disorders.

Smoking, poor diet, excessive alcohol use, and a sedentary lifestyle are all clear contributors to various lifestyles related diseases [22, 23]. According to

research, even tiny adjustments to one's behavior can have a significant impact. Ford et al. [15] found that those who did not smoke, had a body mass index of less than 30 $kg/m_2$, engaged in 3.5 hours of physical activity per week, and consumed a nutritious diet had a 78% decreased risk of getting a chronic illness throughout the course of the 8-year trial. The risks of myocardial infarction, stroke, cancer, and type 2 diabetes all decreased by 93%, 81%, 50%, and 36% respectively. A change in physical activity level alone would result in an increase in life expectancy of between 2.8 and 7.8 years for men and between 4.6 and 7.3 years for women, depending on the degree of the increase in activity, according to actual disease and death rates of physically active and inactive people in Denmark aged 30 to 80 years [16].

Despite this convincing evidence, neither general medical treatment nor modern physiotherapy practice is dominated by lifestyle-related diseases or methods for avoiding, reversing, and managing them [24]. The idea of health is drastically altering in response to these modern health trends and goals [18]. The focus of healthcare is shifting from disease models to health models on a global scale. Contrarily, lifestyle-related diseases are multi-factorial illnesses that are influenced by both environmental and genetic variables and are brought on by the interaction of numerous risk factors [13]. These illnesses have sneaky onsets, a protracted incubation period, and a quick progression. Identifying and treating large numbers of patients in a timely manner is challenging. Additionally, as the majority of lifestyle-related diseases still have unclear etiologies and pathogens

and poor therapeutic outcomes, it is important from a practical standpoint to prevent the development of lifestyle-related diseases.

In terms of the characteristics of lifestyle-related diseases and contemporary health trends, early disease prediction has significant research ramifications. It is one of the key steps in preventing and treating diseases that are caused by a person's lifestyle because identifying population risks prior to the onset of diseases can help people change their lifestyles as soon as possible, especially the life behaviors of high-risk groups, lowering the risk of disease [25]. The primary tool for assessing and preventing lifestyle-related diseases is the disease prediction model [26]. Disease prediction models specifically establish an intelligent model to predict the probability of a specific disease at a specific point in the future, classify high-risk groups in accordance with the probability cut-off point, and conducts behavior, diet, and other interventions to prevent future disease. It can fall under the heading of illness prevention. In other words, the disease prediction model may show assessment subjects about the likelihood that they will become ill in the future and anticipate this likelihood, as well as advise them on how to manage their own health.

### 1.3 Research Status of LRDs Prediction

The original disease prediction model is a disease prediction model of coronary heart disease, which was established by the United States based on the Framingham cohort study [27], and other cardiovascular disease risk assessment models with various markers[28, 29]. The disease prediction models have grad-

ually expanded from cardiovascular disease to include a variety of diseases. For instance, the United States has developed a model for predicting stroke based on the Framing cohort [30]. The Cox proportional hazards model approach is used in this model to create an individual stroke risk model for American whites. Age, systolic blood pressure, hypertension, smoking, atrial fibrillation, left ventricular hypertrophy, and other cardiac conditions were among the factors in the model (i.e., myocardial infarction, congestive heart failure, coronary insufficiency, and intermittent cardiac claudication). Additionally, several nations are actively creating and validating disease prediction models for various diseases appropriate for their particular ethnic characteristics because populations in different countries have varied disease spectrums and prevalence risk factors. For instance, the UK Prospective Diabetes Study (UKPDS) [31], Harvard Cancer Risk Assessment Tool [32], the breast cancer disease prediction-Gail model [33], and a prediction model for lung cancer proposed by the Cancer Research Center of University of Texas Anderson [34].

Machine learning (ML) techniques, a subset of artificial intelligence techniques, employ computer systems to predict diseases using statistical models and algorithms, opening up a wide range of opportunities for illness prevention [25]. Researchers have utilized a number of ML algorithms to predict various diseases in the field of disease prediction. For instance, the use of ensemble techniques for the early diagnosis of coronary heart disease [35]; the use of support vector machines to detect pre-diabetes and diabetes [36]; the use of random forest algorithms to predict the risk of diabetes in the population examined physically [37];

To predict hypertension, a combination of sub type (the least absolute shrinkage and selection operator, LASSO) and support vector machine recursive feature elimination (SVMRFE) was used [38]. A new ensemble learning-based framework for the early detection of type 2 diabetes utilizing lifestyle markers was also developed [39].

Our study employed Web of Science and Google Scholar as search engines to thoroughly analyze the current research status of LRDs prediction. The search was limited to conference and journal papers published between 2013 and 2022. It is important to note that lifestyle-related diseases are a disease set including those diseases related to lifestyles. Since our aim was to investigate the research status on the prediction of LRDs diseases, the most common LRDs diseases (i.e. hypertension, diabetes, obesity, overweight, and coronary heart disease) were represented for analysis. Searches were conducted with terms including lifestyle diseases (this expression was more commonly used in earlier papers), lifestyle-related diseases, hypertension, diabetes, obesity, coronary heart disease (CHD), and cardiovascular disease (CVD).

Specifically, 45 papers are studied. Data extraction included the author's name, year of publication, predicted disease, type of model, and the specific model used. The categories of models were mainly divided into statistical models (SM) and machine learning models (ML). Statistical models are mainly used to discover correlations between variables and thus predict the output, while machine learning models build analytical systems by learning from data and do not rely on explicit rules of construction [40]. Statistical modeling is more about discovering

relationships between variables and the importance of those relationships, without training or testing. Machine learning, on the other hand, aims to obtain models that can make repeatable predictions in order to obtain the best performance on the test set. Therefore the model category as statistical or machine learning models is classified depending on whether the models were trained and tested in the studied papers.

| **Authors** | **Year** | **Diseases** | **Model Category** | **Models** |
|---|---|---|---|---|
| Kumari et al. [41] | 2013 | Diabetes | ML | Support Vector Machine (SVM) |
| Dalakleidi et al. [42] | 2013 | Diabetes | ML | Logistic Regression Model (LRM) and Decision Tree (DT) |
| Ford E S [43] | 2013 | CVD | SM | Framingham |
| Wang et al.[44] | 2014 | Obesity | SM | SVM, k-Nearest Neighbors (kNN), and DT |
| Dugan et al.[45] | 2015 | Obesity | ML | Random Forest (RF), J48, ID3, Naive Bayes Network (NB), and Bayes trained |
| Nai-Arun N et al. [46] | 2015 | Diabetes | ML | DT, Neural Networks, LRM and NB |
| Lingren et al. [47] | 2016 | Obesity | ML | SVM and NB |
| LaFreniere et al. [48] | 2016 | Hypertension | ML | Neural Networks |
| Vartiainen E et al. [49] | 2016 | Cardiovascular diseases | SM | FINRISK Risk Calculator |
| Weng et al. [50] | 2017 | CHD | ML | Neural Networks |
| Montañez et al. [51] | 2017 | Obesity | ML | Gradient Boosting Decision Tree (GBDT), Linear Regression (LR), Regression Trees (RT), KNN, SVM, RF, and MLFFNN |

Continued on next page

| Authors | Year | Diseases | Model Category | Models |
|---|---|---|---|---|
| Rajput et al. [52] | 2018 | Obesity | ML | Neural Networks |
| Ye et al. [53] | 2018 | Hypertension | ML | Extreme Gradient Boosting (XG-Boost) |
| Nour et al. [54] | 2018 | Hypertension | ML | DT and RF |
| Patnaik et al. [55] | 2018 | Hypertension | ML | SVM |
| López-Martínez et al. [56] | 2018 | Hypertension | ML | LRM |
| Effoe V et al. [57] | 2018 | Cardiovascular diseases | SM | Cox Regression |
| Machorro-Cano et al. [58] | 2019 | Obesity | ML | J48 DT |
| Daanouni et al. [59] | 2019 | Diabetes | ML | KNN and DT |
| Ahuja et al. [60] | 2019 | Diabetes | ML | SVM, Multi-Layer Perceptron (MLP), LRM, RF and DT |
| Daanouni et al. [59] | 2019 | Diabetes | ML | Neural Networks |
| Yahyaoui A et al. [61] | 2019 | Diabetes | ML | Neural Networks |
| López-Martínez F et al. [62] | 2020 | Hypertension | ML | Neural Networks |
| Tjahjadi et al. [63] | 2020 | Hypertension | ML | KNN |
| Alpan et al. [64] | 2020 | Diabetes | ML | BN, J48, RF, KNN, and SVM |
| Rahman et al. [65] | 2020 | Diabetes | ML | Neural Networks |
| Memon S A [66] | 2020 | Obesity | ML | L1-regularized regression |
| Singh B [67] | 2020 | Overweight | ML | MLP |
| Shukla AK [68] | 2020 | Diabetes | ML | LRM |
| Islam et al. [69] | 2020 | Diabetes | ML | NB and LRM, RF |
| Abdel-Basset, M et al. [70] | 2020 | Diabetes | ML | SVM, DTs, RF, and LR |
| Aminian A et al. [71] | 2020 | Cardiovascular diseases | ML | RF |

Continued on next page

| Authors | Year | Diseases | Model Category | Models |
|---|---|---|---|---|
| Athanasiou M et al. [72] | 2020 | Cardiovascular diseases | ML | XGBoost |
| Rezaee M et al.[73] | 2020 | Cardiovascular diseases and diabetes | SM | Cox Regression |
| Yaganteeswarudu A et al. [74] | 2020 | Diabetes, Diabetic Retinopathy, Heart Disease, and Breast Cancer | ML | RF, SVM, Neural Networks |
| Chaves L and Marques G [75] | 2021 | Diabetes | ML | Neural Networks |
| Shorewala V [36] | 2021 | coronary heart disease | ML | kNN, LRM and NB |
| Wang K et al. [76] | 2021 | coronary heart disease | SM and ML | Cox regression and XGBoost |
| Islam M M and Shamsuddin R [38] | 2021 | Hypertension | ML | Neural Networks |
| Islam M M et al. [77] | 2021 | Hypertension | ML | SVM |
| Li L et al. [78] | 2021 | Diabetes | SM | Multiple Cox regression |
| Ferdowsy F et al. [79] | 2021 | Obesity | ML | kNN, RF, LRM, MLP, SVM, NB, ADA, DT and GBDT |
| Rashid J et al. [80] | 2022 | Breast cancer, diabetes, heart disease, hepatitis, and kidney disease | ML | Neural Networks |
| Gupta A and Singh A. [81] | 2022 | Heart disease and diabetes | ML | Adaptive Boosting (AdaBoost) |

Continued on next page

| Authors | Year | Diseases | Model Category | Models |
|---|---|---|---|---|
| Yan J et al. [82] | 2022 | coronary heart disease | ML | XGBoost, Light Gradient Boosting Machine (LightGBM), RF, NGBoost, LRM and MLP |

**Table 1.1:** Studied papers about LRDs prediction.

Based on Table 1.1, it observed that there already exist numerous studies focusing on risk prediction of LRDs diseases, among which 39 papers use machine learning-based models for LRDs prediction, 5 papers use statistical based models for LRDs prediction, and 1 paper uses both types of models for analysis. In general, machine learning is increasingly applied in LRDs prediction and is one of the current research hotpots for LRDs prediction.

## 1.4 Problem Statement and Objectives

According to literature studies, almost all existing prediction studies (91%) focus on single disease prediction, with 14 papers focusing on diabetes prediction, 9 papers on hypertension prediction, 9 papers on overweight or obesity prediction, 9 papers on cardiovascular disease prediction, and only 4 studies focusing on multiple disease prediction. Specifically, Yaganteeswarudu [74] proposed a system using the Flask API to predict multiple diseases including diabetes, diabetic retinopathy, heart disease, and breast cancer. This system uses different datasets to train different machine-learning models for different diseases. Rezaee M et al.

[73] achieved consistent discrimination performance for multiple cardiovascular diseases and type-2 diabetes using prediction models derived from Cox proportional risk regression. These models contain multiple shared predictor variables and can be integrated into a single platform to enhance clinical stratification to influence health outcomes. Moreover, Rashid J et al. [80] proposed a new augmented artificial intelligence approach using artificial neural networks (ANN) and particle swarm optimization (PSO) to predict five prevalent chronic diseases including breast cancer, diabetes, heart disease, hepatitis, and kidney disease using five open-source datasets. Further, Gupta A et al. employed genetic algorithm based on recursive feature elimination and AdaBoost to predict two lifestyle diseases (heart disease and diabetes) using two open-source datasets with missing values. On further analysis, the quantitative relationship between models and diseases in the studied papers was as follows.

- One to one: almost all studied papers only focused on predicting a single disease.
- one-to-many: three studied papers used the same model and different datasets to predict multiple diseases.
- many-to-many: only one studied paper employed different models to predict different diseases in different datasets.

Based on the above analysis, existing studies are unable to intelligently identify key features of diseases while building prediction models with different structures and robustness for different LRDs. Therefore, our objective is to design

an intelligent risk prediction framework for LRDs that can smartly identify key features of different LRDs for dirty real medical data, accurately predict the risk of LRDs and visualize prediction results.

## 1.5 Organization of Thesis

The present thesis is organized in 6 chapters as shown in Figure 1.1. Following the introduction in Chapter 1, the rest of the thesis chapters are as follows:

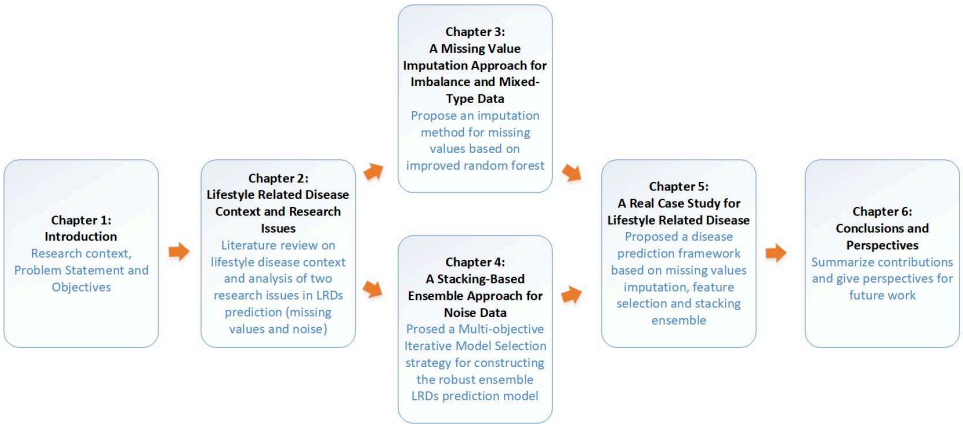



**Figure 1.1:** Chapters organization

In Chapter 2, entitled "Data Characteristics and Proposed LRDs Prediction Framework", introduces the characteristics of the studied health examination data, explains and analyzes technical difficulties of this study, and then introduces the proposed prediction framework for LRDs.

In Chapter 3, entitled "A Missing Value Imputation Approach for Imbalance and Mixed-Type Data", introduces two proposed model including ALWRF and SncALWRFI. Specifically, the structures of ALWRF and SncALWRFI are introduced and Bayesian optimization is employed to optimize their parameters.

Meanwhile, extensive experiments are conducted to evaluate these two models' performance.

In Chapter 4, entitled "A Stacking-Based Ensemble Approach for Noise Data", firstly introduces the proposed Multi-objective Iterative Model Selection (MoItMS) strategy, which use to select individual models for the ensemble model. Meanwhile, ensemble technologies are introduced and the stacking-based ensemble architecture is employed to improve the performance of LRDs risk assessment. Furthermore, extensive testing is performed utilizing real-world data to evaluate the performance of the proposed ensemble model.

In Chapter 5, entitled "A Case Study for Lifestyle Related Disease", the effectiveness of the proposed disease prediction framework is illustrated using a real case in Nanjing, China, taking hypertensive disease as an example.

In Chapter 6, entitled "Conclusions and Perspectives", a brief summary of the main contributions, conclusions, and potential future perspectives is presented.

# Chapter 2. Data Characteristics and Proposed LRDs Prediction Framework

## 2.1 Characteristics of Studied Data

Regular physical examinations have become a crucial component of public health care, contributing to the rapid rise in disease prevention awareness and public health literacy that has led to an enormous increase in physical examination data. In 2020, China's public hospitals and private hospitals performed 179 million and 38 million health examinations, respectively, in which patients with lifestyle-related diseases make up the large majority of those with diseases found by health examination [83]. Meanwhile, people are progressively coming to understand the value of post-examination health services. In order to analyze the risk factors of specific lifestyle-related diseases when conducting health checks, doctors are increasingly focusing on the collection of information about people's lifestyles. Due to the fact that the development of lifestyle-related diseases is closely related to people's unhealthy lifestyle decisions [14].

However, the large amount of physical examination data is now not fully utilized by the majority of medical examination institutions, which leads to data waste and reduces the effectiveness of physical examination. Using efficient and sufficient physical examination data along with artificial intelligence techniques

can more easily and accurately assess each person's physical state [84]. The standard physical examination includes measurements of height, weight, waist circumference, blood pressure, urine, B-ultrasound, and other items. During the physical examination, individuals' age, gender, and other information will be recorded, and some institutions also inquire about their lifestyles. In general, health checks can collect three different types of medical data: administrative, inspection, and lifestyle data [85].

Several studies utilizing health check data to predict LRDs have been conducted. Hui Yang et al. [86] designed an online diabetes risk assessment system and developed an extreme gradient boosting (XGBoost)-based model to predict diabetes risk based on extensive physical examination data. Using data from Japanese health examinations, Mariko Kawasoe et al. [87] constructed a simple and useful clinical prediction model to forecast the 5-year incidence of hypertension in the general Japanese population. Xin Qian et al. [88] developed a cardiovascular disease prediction model using L1 regularized logistic regression with the best predictive performance based on indicators from routine physical examinations. Consequently, health check data is a very good choice for our research. In addition to fully utilizing the ever-growing health data, it may also measure disease risk among individuals and help medical professionals take early preventive action.

Missing values might occur for many causes when collecting health check data. A lot of valuable information is lost when a value is missing, and null values can screw up data mining and produce incorrect results. Moreover, the data from

the health check contains noise and outliers. These values are observations that could be the result of machine or human error, real data, or both. The model's convergence speed and accuracy will be slowed down by noise in the data. Less sensitivity to noisy data will result from increasing the model's robustness. Health check data includes variables that are nominal, binary, and of mixed types. The complexity of data mining will increase as a result of various variable types. In addition, some redundant features in the health check data make the disease prediction model more complex as well. On the other hand, since sick people only make up a portion of all the check-up people, the health check data frequently suffers from an imbalance of positive and negative labels. In addition, as diseases are connected and some become risk factors for others, imbalances in features are frequently present. In conclusion, a variety of characteristics of physical examination data, including incompleteness (missing values), redundant features, noise, mixed types, and imbalance, need to be taken into account in our research.

## 2.2 Research Status of Related Technologies

Based on the analysis in Chapter 1 and the studied data characteristics, three research issues need to be considered in the proposed risk prediction framework for LRDs:

1) Most common prediction techniques are challenging for people to use in accordance with standard processes because medical data that have been collected are dirty and contain a lot of missing values.

2) Effective and precise risk factor identification is essential because removing redundant variables can decrease model complexity and makes it easier to analyze and comprehend model predictions.

3) Since data noise may lower the model's convergence rate and accuracy, it is crucial to research robust models. Enhancing model robustness can reduce sensitivity to noisy data, make models more accurate and offer more reliable auxiliary services.

In conclusion, the proposed risk prediction framework for LRDs must take into account the mentioned three issues: 1) analysis and processing of missing values; 2) identification of key features; and 3) accurate disease prediction. To specifically handle these three issues, the proposed framework must take into account three important techniques.

### 2.2.1 Research Status of Missing Value Processing Methods

As analyzed above, with the construction of modern health information systems, healthcare organizations are experiencing explosive growth in medical data. These medical data contain an abundance of hidden but potentially valuable information, i.e., unknown correlations between diseases and features, and links between diseases with their complications [89]. Such information is useful for medical diagnosis, therapy, and decision-making [90]. However, some unavoidable reasons, such as the early withdrawal of participants from medical research studies and the refusal of participants to attend certain items in medical examinations, can easily result in missing values in research data [91, 92, 93]. Since the

existence of missing values makes it more challenging for people to mine relevant information, many methods for dealing with missing values have been proposed, which can be mainly divided into three categories, namely, deleting missing values, tolerating missing values, and imputing missing values.

Missing value deletion, also known as disregarding missing values, is the process of explicitly deleting instances or variables that contain missing data items to solve the problem of missing data [93]. Although a test pattern with missing values cannot be classified since the deletion procedure would ignore it, deletion methods have the advantage of allowing the normal pattern classification methods to be used directly for complete data [94]. For ignoring missing data, there are two general strategies [95, 94, 93]. First, Listwise Deletion (LD), also known as case-wise deletion, or case removal, is a technique for removing instances (rows, cases) with missing data. This technique is also known as complete case analysis because it only keeps complete cases for analysis (CCA). The analysis is then restricted to those observations for which all values are observed, which frequently leads to biased estimates and loss of precision [18] because this method excludes all cases with missing values for any variable of interest. The second technique is known as Pairwise Deletion (PD) or Available Case Analysis (ACA), also referred to as variable deletion, and it is used to delete variables (columns) with missing data [96]. This method analyzes all situations in which the variables of interest are present, using as much data from each case as is feasible rather than excluding the entire case. Even though some of its variables have missing values, it can nevertheless maintain the most amount of data possible for analysis since it uses

distinct sample sizes for each variable [96]. As a result, the ACA approach has a larger sample size than the CCA method.

In the second type of missing value processing approach, the model is built with some strategies to tolerate missing values. For instance, XGBoost, LightGBM, and Catboost ensemble tree models and decision trees both process missing data during training. These models specifically attempt, during the decision tree construction process, to allocate samples with missing values in the features selected as split points to the left sub-tree or the right sub-tree, and then analyze which side will reduce the loss. This method preserves all data while also assisting in the discovery of hidden information in missing data. Nevertheless, these techniques only work with certain model architectures, which makes the model more complex.

In the third type of missing value processing method, the value estimated by the model is used to replace the missing value. Early approaches for imputing missing data were specifically motivated by traditional statistical models and estimate processes, which are referred to as imputation methods based on statistics. These techniques are designed to model the information included in the non-missing parts of the data set in order to as correctly estimate the missing values as possible [97]. Researchers initially substituted missing values with the mean, median, mode, and zero values. The disadvantage is that when there are numerous missing data, a significant portion of the data is replaced by the same value (i.e., mean, median, mode, zero), which can easily lead to serious deviation. The mean imputation approach should not be used, according to certain recent

research that has demonstrated its shortcomings [98, 99]. The in-depth study on missing values has been accompanied by the proposal of a number of innovative techniques. For instance, the Least Squares (LS) imputation approach is based on the least squares principle to estimate missing values, whereas the hot-deck imputation method predicts missing values by seeking for the nearest neighbor using non-missing information [100].

Further, the researchers used machine learning models to impute missing values. Machine learning-based imputation approaches are complex processes that often include building a predictive model to estimate values that will substitute those missing [101]. The machine learning-based imputation method often involves building a predictive model to predict the values for missing data. Many machine learning-based imputation methods have been proposed recently, and these methods frequently produce good imputation results. Examples of these methods include imputation methods based on decision trees (DT) [102, 103], imputation using multilayer perceptrons [104], imputation using artificial neural networks (ANNs) [105], and imputation using self-organizing maps (SOMs)[106].

The three missing value processing methods and their advantages and disadvantages are shown in Table 2.1.

| Methods | Advantage | Disadvantage | Example |
|---|---|---|---|
| Deletion | Simple | Ignore valuable information | PD, LD |
| Toleration | Learning hidden information | Increase model complexity; specific model structure | DT, XGBoost |
| Imputation | Independent of predictive models | Additional computing space and time | Mean, KNNI |

**Table 2.1:** The three missing value processing methods and their advantages and disadvantages.

### 2.2.2 Research Status of Feature Selection Methods

As it can be challenging for people to distinguish between significant and superfluous features when gathering data, feature selection is an essential component of data reprocessing. Specifically, feature selection refers to choosing a task-related feature subset from the full set of features in order to reduce the amount of data that must be stored, shorten the time needed to train machine learning models, and enhance the predictive skills of machine learning models. Therefore, feature selection can assist in both the identification of essential features and the elimination of superfluous features. Data mining techniques based on machine learning techniques were used to select the primary characteristics of lifestyle-related diseases. The benefit of this approach is that the outcomes are generated by data analysis without the need for human interaction. This approach is appropriate for those without strong expertise in medicine and uses sophisticated algorithms to guide people in choosing essential factors. Our research belongs to the category of supervised learning because it focuses on the prediction of LRDs disease. We, therefore, concentrate on feature selection for

supervised issues in this study. Three categories of feature selection techniques can be distinguished based on the form of the feature selection [107]:

- Filter: Determine thresholds or the maximum number of features to be selected, and then rank each feature according to specific statistical indicators.
- Wrapper: When choosing alternative feature subsets for the model's training, consider the impact of cross-validation as the optimization objective. Then, choose the best combination.
- Embedded: After the model has been trained, many machine learning models allow for the evaluation of the contribution of each feature to the prediction result. The threshold, or the number of thresholds to be selected, can then be set in accordance with the contribution, and the feature can be chosen.

Three feature selection categories and their advantages and disadvantages are shown in Table 2.2.

| **Category** | **Advantage** | **Disadvantage** | **Example** |
|---|---|---|---|
| Filter | High computational efficiency | Ignore combination effect between features | Pearson correlation coefficient, chi-square test |
| Wrapper | Oriented to algorithm optimization | High complexity and easy to overfit with small samples | Complete search, random search |
| Embedded | Automatically selects features | Need to select loss functions and adjust parameter | Feature Selection Method Based on Tree Model |

**Table 2.2:** Three feature selection categories and their advantages and disadvantages.

### 2.2.3 Research Status of Disease Prediction Methods

Nowadays, a lot of academics are researching disease prediction models and have developed a number of useful models. In earlier research, we investigated the state of various prevalent lifestyle diseases prediction methods (hypertension, diabetes, obesity, overweight, and coronary heart disease). There are specifically 3 statistics-based models: Framingham, FINRISK Risk Calculator, and Cox Regression. The Framingham risk score can be used to calculate a person's 10-year cardiovascular risk, even in those without a history of heart disease. Based on the findings of the Framingham Heart Study, this risk score has been developed. Based on risk factor information and incidence tracking from researchers in the five-year FINRISK study, the FINRISK calculator was developed. Each risk factor that was taken into account while developing the risk coefficients was first evaluated for its impact on disease prevalence and mortality using multivariate analysis. For analyzing the relationship between patient survival time and one or more predictor factors, the Cox Regression model is frequently employed in medical research.

On the other hand, a wide range of machine learning models, including but not limited to SVM, NB, and Neural Networks, have been employed to predict LRDs. These models use the rules to forecast unknown data after automatically analyzing the data. Different machine learning methods are suitable for different types of data [91]. For instance, Although BN does not have severe limitations on the number of samples as well as a high classifier efficiency, the prediction

performance is poor due to the assumed prior model in some situations; the KNN model is challenging people to apply to high-dimensional and sparse data; the SVM model is also simple to manage when the number of sample features and the number of samples are close together. Two different categories of feature selection methods together with their benefits and drawbacks are shown in Table 2.3.

| **Category** | **Advantage** | **Disadvantage** | **Example** |
|---|---|---|---|
| Traditional Statistical Methods | Strong model interpretability | Modeling is based on multiple assumptions; underperform in complex data | Framingham, Cox Regression |
| Machine learning method | High flexibility and learning capability | High model complexity; Poor interpretability | XGBoost, Neural Network |

**Table 2.3:** Two types of disease prediction methods and their advantages and disadvantages

## 2.3 Technical Challenges

As we have already mentioned, as living standards have increased, people's concern for their personal health has increased. To lower risks or postpone the development of contracting lifestyle-related diseases, people have chosen a variety of strategies, including health screenings, diet, and exercise. Over time, a large amount of health and medical information is recorded and stored in detail by the medical information system. The foundation for research on lifestyle-related diseases has been set in this situation by enough health examination data and some lifestyle-related data. Researchers are now concentrating on applying machine learning techniques to mine valuable information hidden in health test data to

assist people in predicting and preventing diseases connected to lifestyle choices. But when data mining is used, two key aspects in gathered medical examination data—missing values and noise—present technical difficulties for the analysis and prediction of lifestyle-related diseases. Next, we will provide an in-depth analysis and introduction of missing values and noise in the dataset of lifestyle-related diseases.

### 2.3.1 Missing Values in Imbalanced and Mixed-type Features

A simple and easy-to-operate missing value processing technique is missing value deletion, but this technique is prone to losing valuable information and is unable to be utilized with data that has a lot of missing values. Furthermore, some predictive models develop techniques to deal with missing values, which can help preserve more useful information but makes the predictive model more complex and only works with specific model structures. The missing value imputation method can keep more valuable information, is more flexible, and is not dependent on the prediction model.

Numerous methods are available in the literature to impute missing values in metrically scaled data, such as imputation by mean, hot-deck [108], k-Nearest Neighbors (kNN) [109], Decision Tree (DT) [110] and Random Forest (RF) [111]. The two types of mean imputation are conditional and unconditional mean imputation, both of which are quick but may destroy the data distribution [99]. The kNN technique finds the k-nearest records to fill in missing values. The kNN strategy has the advantage of simplicity, but it requires searching the entire dataset

to locate the k-nearest neighbors. In addition, as kNN ignores the correlation between covariates, Shahla and Gerhard [112] proposed a sophisticated imputation method for mixed-type data that uses non-parametric nearest-neighbor and takes into account the correlation between covariates. Although it yields smaller imputation errors and higher performance in datasets with significant covariate correlation, it easily encounters disaster in time and space in large-scale datasets since it needs to multiple search datasets and calculates distances between records. Further, researchers prefer tree-based imputation methods like the decision tree and random forest model because of their high interpretability, quick prediction speed, and adaptability for mixed-type datasets. For example, Rahman and Islam [113] employed decision trees and decision forests to impute missing values by dividing and merging records and achieved outperformed results on nine public datasets. Even though they used tree-based approaches to impute missing data, their methods are computationally complicated and demand a lot of memory when merging records from many trees with various structures. In another tree-based example, Nikfalazar and Yeh et al. [114] introduced a new missing value imputation approach that considers mixed-type data by combining decision trees and fuzzy C-means (FCM) [115] with iterative learning. But single decision tree is susceptible to noise, and it is time-consuming to search for the number of clusters and perform clustering.

In the medical field, large-scale datasets with mixed type and imbalance characteristics are widespread [116, 117]. Although advanced methods can reduce imputation errors and improve the quality of missing data, existing methods can-

not perform missing values well in data with mixed types and unbalanced characteristics. As a result, we proposed a new missing value imputation method based on the Adaptive Laplacian Weight Random Forest (ALWRF) and the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC), which can adjust the weight of features adaptively when building a random forest and improve prediction accuracy for imbalance features. As far as we know, our work is the first imputation method to consider both adaptive weights and imbalanced problems based on a tree model. We will give a detailed introduction to the proposed missing value imputation method in Chapter 3.

### 2.3.2 Diverse Noises in Lifestyle Related Disease Context

Data noise is the term for errors or unusual data present in the data. The processing and analysis of data sets can be significantly impacted by these data noises. To discover a suitable approach to deal with data noise, it is required to identify the different types of noise in the data. Two categories of noise—attribute noise and class noise—are generally separated in terms of disease prediction [118]. Class noise happens when examples are incorrectly classified into a class, and attribute noise influences the attribute values of examples in the dataset. Both attribute noise and noise-like noise can affect the classifier's performance [119].In medical data classification, noise can come from multiple sources:

- Human error. Errors in the labeling process, which are more likely to occur in jobs dealing with complex data, can occur due to fatigue, routine, checking each case quickly, or time pressure. In addition, subjectivity also

creates category noise. For example, when there are discrepancies in the labels of multiple experts.

- Machine errors. When machines are responsible for providing automated data, design errors or transient errors can result in incorrect attributes and labels.

- Digitization and filing errors. When creating digital records of inspection cases, categories can be entered incorrectly due to simple mistakes. The same happens when using history.

In particular, noise is a combination of attribute noise and class noise in the medical data of lifestyle-related diseases, where attribute noise is mostly made up of abnormal attribute values, or outliers, which are distinct individual points from the system as a whole. However, these points cannot be ignored because they might potentially have useful information. For instance, a study [120] on diabetes discovered that outliers were indicated by significant disparities between the maximum value of two key characteristics, triglycerides (TG) and low-density lipoprotein (LDL), and the third quartile. Because they belonged to valid patients, these outliers were not excluded. On the other hand, class noise is known as wrong instance labels. In practice, the diagnosis of lifestyle-related diseases is prone to mislabeling. For example, if you speak during blood pressure measurement, blood pressure will increase by 5-19 mmHg; or when blood pressure is measured in a cold environment, blood pressure may increase by 5-23 mmHg; these situations can affect blood pressure measurement and result in a

misdiagnosis. Furthermore, it is impossible for medical professionals to ensure the utmost accuracy of their diagnosis results when making medical diagnoses in the face of complicated lifestyle-related diseases, such as coronary heart disease and tumors.

According to the above analysis, medical data of lifestyle-related diseases inevitably have attribute noises and label noises. In general, there have been two basic strategies for dealing with noisy data in medical data:

- Algorithmic-level methods. These techniques are characterized by being less affected by noisy data. For example, C4.5 [121] uses a pruning strategy to reduce the chance of the tree overfitting due to noise [122].
- Data-level approach. The most well-known type of method in this group is the noise filter [123]. They identify noisy examples, which can be eliminated from the training data.

Using data-level methods to directly delete outliers is easy to lose effective information because there are some in the dataset of lifestyle-related diseases examined that correspond to valid patients. The ensemble approach, an algorithm-level technique, is a great way to lower variation, bias, and noise, and it can combine several individual models as a whole to outperform each individual model. Therefore, in order to improve the accuracy of lifestyle-related disease prediction, we employed an ensemble method in our research to develop a robust disease prediction model, which makes the model less sensitive to noise

and improves the prediction accuracy of LRDs. In Chapter 4, a comprehensive introduction to the proposed ensemble method is presented.

## 2.4 The Overview of Prediction Framework

A framework for LRDs prediction is proposed based on the above findings. The missing value module, feature selection module, and disease prediction module are the three key components of this framework. The method of combining deletion and imputation is chosen as the primary strategy for missing value processing for the significant number of missing values in the data set gathered from lifestyle-related diseases first. The feature selection module employs machine learning-based feature selection to discover key features for lifestyle-related diseases since different lifestyle-related diseases have distinct important features. In order to create a strong ensemble prediction model for lifestyle-related diseases and achieve a more accurate prediction of lifestyle-related diseases, the data processed by the missing value module and the feature selection module are used as the input of the prediction model. Figure 2.1 is a diagram of the proposed prediction framework for LRDs.

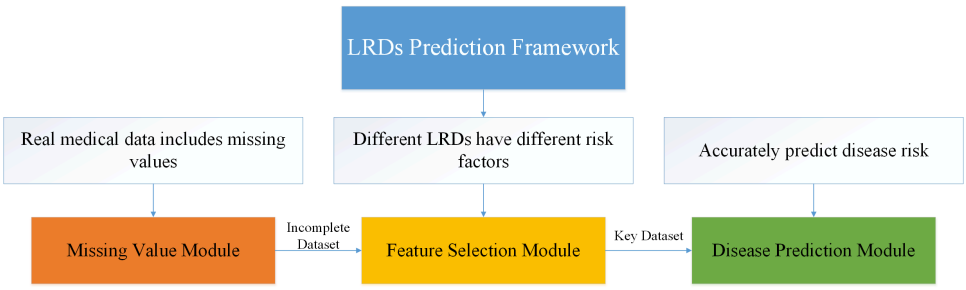



**Figure 2.1:** LRDs prediction framework

### 2.4.1 Missing Value Module

In the missing value module, in order to enable the comprehensive analysis of missing values, missing rates, and missing patterns are visually presented. Firstly, The missing rate analysis help to rapidly comprehend the missing conditions in the data set. Meanwhile, it can also use this information to help choose the processing strategy for missing values. For clarity of definitions, we assume that data set $X$ includes $n$ instances and $k$ features. Let $M$ represent a missing value matrix, where $m_{ij}$ has a value of 0 if any value $x_{ij}(i \leq n, j \leq k)$ in $X$ is observed and 1 otherwise. The total missing rate $MR$ can be represented as

$$MR = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}}{m \times n} \tag{2.1}$$

On the other hand, the missing rate of $i$th row (denoted by $r_i$) can be calculated by

$$MR_{r_i} = \frac{\sum_{j=1}^{k} m_{ij}}{m} \tag{2.2}$$

Finally, the missing rate of $j$th column (denoted by $c_j$) can be computed by

$$MR_{c_j} = \frac{\sum_{i=1}^{n} m_{ij}}{m} \tag{2.3}$$

Second, by displaying the distribution of missing values, such as univariate, monotone, and non-monotone [93], the study of missing patterns can assess how complex missing values are. Finally, by examining the relationships between different

features with missing values, such as MCAR, MAR, and NMAR [95], the study of the missing mechanism can investigate the causes of missing values.

In reality, some features or instances will have a disproportionate number of missing values for a variety of reasons; for instance, 99% of the values will be absent. The major features of lifestyle-related diseases are used in our study to build excellent predictive models, so when features or instances have a large number of missing values, this is difficult to apply in our study. Instead, we will prefer to use the deletion method rather than filling in a large number of estimates. We need to describe the criteria for deleting missing values, or the threshold for using it, in more detail. According to the 80% rule [124], which states that a substance should be removed if its non-missing portion is less than 80% of the sample size as a whole, the suggested prediction framework excludes features or instances whose missing rate is more than 80%. At the same time, the framework provides an interface for customizing the threshold, making it simple for knowledgeable specialists to adjust the threshold based on their own expertise.

There are still some missing values in the dataset even though some features and instances are compelled to be removed in accordance with the threshold setting of the missing rate. The reasons and ways of missing are typically dispersed among several features and instances, making it difficult to simply eliminate them using a deletion procedure. Therefore, to appropriately handle missing values, we shall employ more sophisticated techniques. We suggest a missing value imputation technique in Chapter 3 that can be used with datasets that are imbalanced or mixed types. We employed the proposed missing value imputation

approach as our default missing value handling method in the missing value imputation step since features with characteristics of unbalanced and mixed types are common in datasets of lifestyle-related diseases. Similarly, we incorporate various well-known and excellent imputation methods for missing values, such as MissForest and KNNI, as alternatives or benchmarks in order to provide people with more options.

### 2.4.2 Feature Selection Module

Data mining techniques based on machine learning techniques are employed to select the primary characteristics of lifestyle-related diseases. The benefit of this approach is that the outcomes are generated by data analysis without the need for human interaction. This approach is appropriate for those without strong expertise in medicine and uses sophisticated algorithms to guide people in choosing essential factors. In previous studies, we surveyed existing feature selection methods, and each method has its own advantages and disadvantages. The feature selection of the wrapper has high computation complexity, and the filtering mechanism ignores the connection between the feature and the target variable. As a result, the tree-based strategy in the embedding method is employed for feature selection in the proposed prediction framework. Splitting into tree-based approaches occurs in the classification model due to Gini impurity or information gain/entropy, whereas it occurs in the regression model due to variance. Using techniques like random forests and gradient boosting, features are chosen according to the relevance of each one. Generally, features with high im-

portance are more likely to have an impact on the target feature. The proposed prediction framework uses the random forest importance approach as the main algorithm of the feature selection module because the random forest has high generalization capabilities and is appropriate for large-scale datasets.

Specifically, the random forest feature importance evaluation calculates the mean value of each feature's contribution to each tree in the random forest. There are two techniques to obtain the final collection of key features after assessing the importance of each feature: 1) select Top-N features, 2) Select larger than the set threshold. Since the value of N is difficult to determine and in order to keep as many task-related features as possible, the feature selection module selects according to the important threshold of the feature.

### 2.4.3 Disease Prediction Module

As we previously mentioned, a variety of machine learning algorithms have been utilized by researchers to estimate the risk of various diseases in the field of disease risk prediction. There are various noises in the data set, which threaten the accuracy of the disease prediction model. Therefore, in order to build a robust prediction model for LRDs, we will employ ensemble techniques to reduce the impact of noise. We propose a stacked ensemble method in Chapter 4, a technique that can be used on datasets with diverse noise. We adopted the proposed stacked ensemble method as the default prediction method for the disease prediction module.

Specifically, the disease prediction module includes visualization of the development of forecasting models, evaluation of the models, and interpretation of forecasting results. Visualization of model development can better explain the prediction process of lifestyle-related diseases. The model's evaluation is also crucial because it defines how usable the final model will be. The evaluation index provides a quantitative index of the quality of the algorithm or parameters and is designed to input the same data into several algorithm models or the same algorithm model with varied parameters. It is frequently important to employ a variety of various indications while evaluating a model. The majority of the numerous evaluation indicators can only indicate a portion of the model's performance. If the evaluation indicators are not used properly, flaws with the model itself cannot be detected, which will result in incorrect inferences. The interpretation of prediction results can provide people with rich information.

# Chapter 3. A Missing Value Imputation Approach for Imbalance and Mixed-Type Data

## 3.1 Methodology of the Proposed Imputation Method

### 3.1.1 Adaptive Laplacian Weight Random Forest (ALWRF) Method

On mixed-type data, tree-based models have a natural advantage because their construction is concentrated only on the information gain of features rather than the distance between cases [93]. On the other hand, tree-based models show high interpretability compared to algorithms such as neural networks, because their routes from the root node to the leaf node represent a rule [125]. A decision tree is one of the most representative tree-based models. The decision tree starts from the root node of the tree, continually splits by selecting the optimal attribute, and builds the tree nodes one by one until a stopping condition of tree building is satisfied. There are two typical stopping conditions, including no samples in the child nodes and exhaustion of attributes. As a single decision tree frequently suffers from overfitting, ensemble approaches based on decision trees have been proposed including Boosting[126] and Bagging [127]. Random forest is an ensemble algorithm based on the bagging approach that has strong anti-noise properties and can perform effectively on large data sets [110, 128].

Meanwhile, as it blends the idea of the ensemble with randomization, overfitting is well-controlled. In particular, random forest uses the bootstrap technique to randomly draw samples from original samples to build a single decision tree, and then repeat this process a specific number of times (the number of trees) [125]. Finally, the final prediction result is obtained by combining these decision trees.

In random forests, features with high quality are not fully used because features are selected consistently and randomly to construct a feature subspace. As a result, the random forest's performance may be limited, because all features, including those with little or no information, have the same probability [129]. From the standpoint of feature subspace selection, some better random forest methods have been developed. Amaratunga and Cabrera et al. [130] proposed enriched random forests: choose the eligible subsets at each node by weighted random sampling instead of simple random sampling, with the weights tilted in favor of the informative features. Then, stratified Random Forests [131] utilized the weights that obtained by Fisher discriminant projection to divide the features into two parts, namely strong and weak features. However, it needs to determine the segmentation threshold of strong and weak features, as well as the amount of strong and weak characteristics. Further, Liang and Huang et al. [132] took advantage of the Laplacian score [133] to quantify the importance of different features by considering their locality preserving power and then generated a set of diverse subspaces by weighted random sampling. To sum up, these studies are mostly concerned with estimating features and raising the weight of excellent features. However, the diversity of random forests is easily reduced by utilizing fixed

weights. To improve this situation, we proposed an adaptive Laplacian weight random forest (ALWRF) by dynamically adjusting the weight when constructing trees.

As the decision tree is the basic model of random forest, common decision tree algorithms are introduced first, ie., ID3[134], C4.5 [135], Classification and Regression Tree (CART) [136]. The ID3 algorithm iterates through every unused attribute and calculates the entropy or the information gain of that attribute and it then selects the attribute which has the smallest entropy (or largest information gain) value. ID3 is harder to use on continuous data than on factored data (factored data has a discrete number of possible values, thus reducing the possible branch points) [135]. The C4.5 algorithm is an extension of the earlier ID3 algorithm and can be used for classification. ID3 and C4.5 are time-consuming because of logarithmic operations in entropy models. In the CART algorithm, each node has less than or equal to two children. The bisection method can simplify the scale of decision trees and improve the efficiency of generating decision trees. On the other hand, the CART algorithm can be used to create both classification trees and regression trees, which is suitable for categorical missing values and numerical missing values [114]. Therefore, the CART algorithm is employed as a basic model in ALWRF. In the CART algorithm, the outputs for the classification tree and regression tree are discrete value and continuous value respectively. In detail, the output in the classification tree is the majority class of the leaf node, while the regression tree uses the mean value of the leaf node as the output. In addition, the CART algorithm uses the Gini coefficient as the

impurity of variables, which can reduce the complexity of logarithmic operations compared with ID3 and C4.5. The smaller Gini coefficient shows the feature is better. The equation for the Gini coefficient is:

$$Gini(D) = \sum_{i=1}^{n} p(x_i) * (1 - p(x_i)) = 1 - \sum_{i=1}^{n} p(x_i^2) \tag{3.1}$$

where $p(x_i)$ is the probability of occurrence of category $x_i$ and $n$ is the number of categories. $Gini(D)$ reflects the probability of two randomly drawn samples from dataset $D$ whose class labels are inconsistent. Therefore, the smaller $Gini(D)$ represents the higher purity of the dataset $D$.

In our work, in order to evaluate the feature importance for enhancing the performance of random forest, we resort to the adaptive feature selection technique termed adaptive Laplacian score. The Laplace score is a classical and popular feature selection algorithm in filter style, which aims to find the most discriminative features [133]. To avoid confusion, we assume that training data with $n$ samples and $d$ dimension. its data matrix can be represented as $X \in R^{n \times d}$. Each row in $X = (x_1, x_2, ..., x_n)^T$ corresponds to a sample, while each column corresponds to a feature. $x_i \in R^d$ is the $i$-th sample. Thereby, the data matrix can also be denoted as $X = (f_1, f_2, ..., f_d)$, where $f_j \in R_n$ is the $j$-th feature. Particularly, a k-nearest neighbor graph is first constructed, which is used to calculate the Laplacian scores of different features by considering their locality-preserving power. We denote this graph as $G = \{V, E\}$, where $V = \{x_1, x_2, ..., x_n\}$ is the set of the training samples and $E \in R^{nn}$ is the adjacent matrix. Here, we use the 5-nearest neighbor and the cosine similarity to compute the edge weights.

That is

$$E = \{x_{ij}\}_{n \times n} \tag{3.2}$$

$$e_{ij} = \begin{cases} \dfrac{\psi(x_i, x_j)}{\sqrt{\psi(x_i, x_i).\, \psi(x_j, x_j)}}, & x_i \in kNN(x_j) \text{ or } x_j \in kNN(x_i), \\ 0 \quad , & otherwise, \end{cases} \tag{3.3}$$

where $\psi(.)$ computes the inner product of two vectors and $kNN(x_i)$ denotes the set of k-nearest neighbors of $x_i$. Let $D \in R^{nn}$ be the degree matrix, which is a diagonal matrix with its $(i, i)$-th element being the sum of the $i$-th row in $E$. Let the graph Laplacian of $G$ be denoted as $L = D - E$. Then, the Laplacian score of the $i$-th feature $f_i$ can be computed as

$$s_i = \frac{\widetilde{f_i}^\top L \widetilde{f_i}}{\widetilde{f_i}^\top D \widetilde{f_i}} \tag{3.4}$$

$$\widetilde{f_i} = f_i - \frac{\widetilde{f_i}^\top De}{e^\top De} e \tag{3.5}$$

where $e = (1, ..., 1)^\top$. According to equation (3.4), all Laplacian scores of $d$ features can be denoted as $S = (s_1, s_2, ..., s_d)^\top$. As a smaller Laplacian score indicates that this feature can better preserve the locality information and therefore can be viewed as a feature of greater importance, the feature weight of the feature $f_i$ can be defined as $\ell_i = 1 - s_i$. Then normalized feature weights can be denote

as $\widetilde{L} = (\widetilde{\ell}_1, \widetilde{\ell}_2, ..., \widetilde{\ell}_d)$. The $\widetilde{\ell}_i$ is computed by

$$\widetilde{\ell}_i = \frac{\ell_i}{\sum_{j=1}^{d} \ell_j} \tag{3.6}$$

The computed weights $\widetilde{L}$ serve as an initial indicator of the importance of each feature and then diversified random subspaces are generated using the weighted random sampling. With the construction of the Laplacian-weighted random forest, the weights of features are adjusted according to the importance of features on prediction. In detail, the importance of features on prediction can be estimated by the accuracy of out-of-bag (OOB) data after adding random noise in the process of constructing a random forest. Generally, the higher importance of a feature on prediction means that changing its value makes predictions more prone to errors. Specifically, the importance of features on prediction in random forests is the sum of importance in all decision trees. We assume that the number of trees in the random forest is $m$, and the already established set of decision trees is $T = \{t_1, t_2, ..., t_m\}$. The importance of $i\text{-}th$ feature on prediction is calculated by

$$\iota_i = \frac{\sum_{j=1}^{m} e_{i,j}^{OOB1} - e_{i,j}^{OOB2}}{m} \tag{3.7}$$

where $e_{i,j}^{OOB1}$ is the error of corresponding out-of-bag data in $j\text{-}th$ decision tree and $e_{i,j}^{OOB2}$ is the error of out-of-bag data with randomly added noise. Similarly, the importance of features on prediction can be normalized and denoted as $\widetilde{I} = \{\widetilde{\iota}_1, \widetilde{\iota}_2, ..., \widetilde{\iota}_d\}$

With the increase of the decision tree, the weights of features are adjusted by $\widetilde{I}$. The adaptive weights of features are computed by

$$\widetilde{w}_i = \frac{(1-\mu) \times \widetilde{\ell}_i + \mu \times \widetilde{\iota}_i}{2} \tag{3.8}$$

where $\widetilde{\ell}_i$ is the normalized Laplacian weight for the $i\text{-}th$ feature and $\widetilde{\iota}_i$ is the normalized importance of the $i\text{-}th$ features on prediction. $\mu$ is an adjusted parameter which is the ratio of the number of trees that have been constructed to the number of trees that needs to be generated. The interval for updating weights is $\gamma$ which means features' weights are updated in specific iterative times. Additionally, a random operator $\epsilon$ is employed to increase the diversity of trees. Specifically, the weights of features are updated according to a frequency that they are selected when building decision trees. Therefore, the selection probabilities of features with lower weight are increased, which helps construct various trees. Assuming that the number of selected times for features is $N = (\nu_1, \nu_2, ..., \nu_d)$ in decision trees that have been built, and the selected probability of $i\text{-}th$ feature is defined as

$$\rho_i = \frac{\nu_i}{\sum_{j=1}^{d} \nu_j} \tag{3.9}$$

As the smaller number of selected times shows the higher locality and lower importance for the feature, $\epsilon$ is set to 0.9. When the random number is larger than $\epsilon$, the weight of $i\text{-}th$ feature can be updated by

$$\widetilde{w}_i' = \frac{\widetilde{w}_i + \rho_i}{2} \tag{3.10}$$

Based on the previous introduction, the proposed adaptive Laplacian weighted random forest (ALWRF) is shown in algorithm 1.

**Algorithm 1** The adaptive Laplacian weighted random forest (ALWRF)

**Input:** $D$: A data set with $n$ rows and $d$ columns;

$m$: The number of trees;

$\gamma$: The interval to update the weights

**Output:** $ALWRF$

$\widetilde{L} = \{\widetilde{\ell}_1, \widetilde{\ell}_2, ..., \widetilde{\ell}_d\} \leftarrow$ The normalized Laplacian weighted

**for** $i = 0$ *to* $m$ **do**

$V = \{\nu_1, \nu_2, ...\nu_d\} \leftarrow$ The number of selected times for features

$DT \leftarrow \emptyset$

**while** *True* **do**

**if** $DT$ *meets conditions* **then**

break

**end**

$\widetilde{D}^i \leftarrow$ Sampling $m$ times with replacement from $D$

$\widetilde{D}^i_{oob} \leftarrow$ The Corresponding out-of-bag data

$W_c \leftarrow \widetilde{W}$

$random \leftarrow$ A random number in the range (0,1)

**if** $random > \epsilon$ **then**

$W_c^{'} \leftarrow$ Update weights by equation (3.10)

$F_{sub} \leftarrow$ Weighted random sampling of feature subsets using $W_c^{'}$

**else**

$F_{sub} \leftarrow$ Weighted random sampling of feature subsets using $W_c$

**end**

$f_j \leftarrow$ Select the optimal splitting feature using $\widetilde{D}_i$

$DT \leftarrow$ Generate branches and update $DT$

$\nu_j \leftarrow \nu_j + 1$

**end**

$ALWRF \leftarrow ALWRF \cup DT$

**if** $len(ALWRF) \bmod \gamma$ *equals* $0$ **then**

$\widetilde{I} = \{\widetilde{\iota}_1, \widetilde{\iota}_2, ..., \widetilde{\iota}_d\} \leftarrow$ Calculate the normalized importance of features using $D_{oob}$

$\widetilde{W} \leftarrow$ Update weights by equation (3.8)

**end**

**end**

On the other hand, the training and testing process for the proposed ALWRF is similar to a random forest. Five steps involved in the ALWRF:

Step 1: The dataset($n \times d$) is divided into training data ($n_1 \times d$) and testing data ($n_2 \times d$), where $n_1 + n_2 = n$.

Step 2: In ALWRF $n_1$ number of random records are taken from the training data set having $d$ number of records.

Step 3: Individual decision trees are constructed for each sample based on adaptive Laplacian weights.

Step 4: Each decision tree will generate an output.

Step 5: Final output is considered based on majority Voting or averaging for classification and regression respectively using the testing data.

The training and testing process for adaptive Laplacian weighted random forest is shown in Figure 3.1.

### 3.1.2 Oversampling Technique: SMOTE-NC

Imbalanced classifications pose a challenge for missing value imputation algorithms used for classification were designed around the assumption of an equal number of samples for each class. This results in algorithms that have poor predictive performance, specifically for the minority class[137]. Many nominal features with missing values have an imbalanced class distribution in medical data. For example, when diabetes is a feature to predict hypertension, the class of diabetes is the majority and the class of health is a minority. Therefore, imputation algorithms have to pay more attention to incomplete and imbalanced

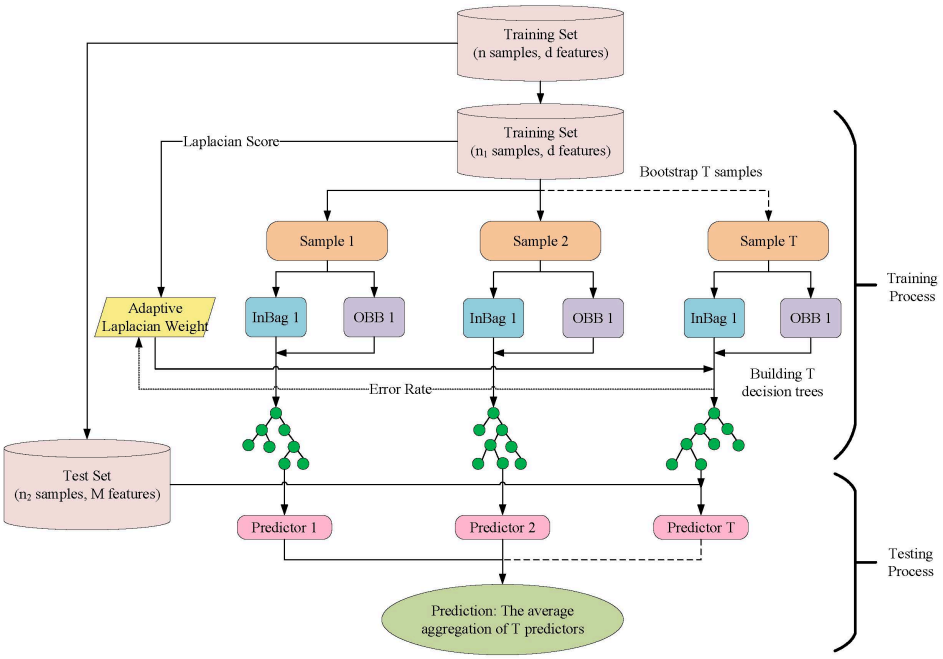



**Figure 3.1:** The training and testing process for adaptive Laplacian weighted random forest

features. Existing methods for solving the problem of imbalanced data mainly focus on the algorithm level [138] and the data level [137]. At the algorithm level, mainly combined with the characteristics of imbalanced data, to improve the accuracy of minority samples [90]. Although this method retains the original data distribution, its usual range is relatively limited [139]. At the data level, the imbalanced level of data is reduced or eliminated mainly by changing the sample distribution of data. Common approaches at the data level contain oversampling of the minority class or undersampling of the majority class. Undersampling technologies have the risk of losing important concepts because they remove a part of the data from the majority classes. At the same time, when the number

of observed data is small, undersampling produces smaller data sets, which may limit the performance of models.

Although the random forest method uses the ensemble idea to preserve the original data distribution and improve the performance of a single decision tree in imbalanced data, its application is limited in highly unbalanced data [139]. In this work, an oversampling technique for mixed-type data is employed to overcome the imbalanced problem. Random oversampling and Synthetic Minority Over-sampling TEchnique (SMOTE) [140] are two popular oversampling methods. Random oversampling reduces data imbalance by randomly copying minority samples, but blind copying may lead to overfitting [141]. The SMOTE algorithm uses linear interpolation to synthesize a new minority sample between some minority samples, which effectively alleviates the risk of overfitting. Although more improved SMOTE methods have been proposed [142, 90, 137], they introduce more computations and parameters. For example, Last and Douzas et al. [137] proposed an advanced oversampling method combining K-Means [143] and SMOTE, which avoids the generation of noise and effectively overcoming the imbalance between classes and within classes. However, this method introduces additional clustering calculations and additional parameters (i.e., the number of clusters $k$ and the density $de$) compared to the naive SMOTE method. Therefore, we resort Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC) [144] to improve imputation performance when facing incomplete and imbalanced features, which creates synthetic data for categorical as well as quantitative features in the data set. The steps of the SMOTE-NC

algorithm are described below and an example of nearest neighbor computation for SMOTE-NC is demonstrated in Table 3.1. Here, $Med^2$ is the median of the standard deviations of continuous features of the minority class.

Step 1: Median calculation. Calculate the median of the standard deviations of all continuous features of the minority class. If the nominal features differ between a sample and its potential nearest neighbors, then this median is included in the Euclidean distance computation. The median is used to penalize the variance of nominal features, the amount of which is related to the typical variance of continuous feature values.

Step 2. Nearest neighbor calculation. Calculate the Euclidean distance between the feature vector that is identifying the k-nearest neighbors (minority class samples) and other feature vectors (minority class samples) using a continuous feature space. For each distinct nominal feature between the considered feature vector and its potential nearest neighbor, including the median of the standard deviations previously computed, in the Euclidean distance computation.

Step 3. Populate the synthetic sample. The continuous features of the new synthetic minority class sample are created using the same approach of SMOTE [140] as described earlier. The nominal feature is given the value occurring in the majority of the k-nearest neighbors (mode).

| | |
|---|---|
| **Two Cases** | F1 = {1 2 3 A B C}, F2 = {4 6 5 A D E} |
| **Median Calculation** | It includes twice for the 5th feature: B→D and the 6th: C→E, which differ for the two feature vectors. |
| **Nearest Neighbor Calculation** | Euclidean Distance: $sqrt[(4-1)^2+(6-2)^2+(5-3)^2+Med^2+Med^2]$ |

**Table 3.1:** Example of nearest neighbor computation for SMOTE-NC.

### 3.1.3 The Proposed Imputation Method

The random forest method is suitable for imputing incomplete and mixed-type data as it works for classification and regression tasks[111]. We apply the proposed adaptive Laplacian weight random forest and the SMOTE-NC method to impute incomplete data with the characteristics of imbalance and mixed type, called SMOTE-NC and ALWRF Imputation (SncALWRFI). Specifically, its procedure is iterative, in which it uses mean and mode values to replace missing data and then it updates missing values on each successive iteration. Consider a given dataset $D$, where The feature set is $F$. The features can be either numerical or categorical. The SncALWRFI method includes 6 steps as follows:

Step 1. Calculate the missing rate of all features $F$ with missing values, and sort the features in descending order. The sorted feature set is donated as $\widetilde{F}(\widetilde{F} \subseteq F)$.

Step 2. Calculate an indicator matrix (donate as $M$) to record the location of missing values, where observed values are 1 and missing values are 0. Then the average of the numerical features and the mode of the categorical features are used to initially impute missing values, donate as $D^{'}$.

Step 3. For each feature $f_i \in \widetilde{F}$ that has a missing value for some of the records, the full dataset $D^{'}$ is divided into two subsets $D_I^i$ and $D_C^i$ according to the indicator matrix $M$, where $D_I^i$ contains all records with missing values at the feature $f_i$ and $D_C^i$ contains records with no missing value at the feature $f_i$.

Step 4. Some available values (value = 1) in the data matrix are set to missing (value = 0) and then these values will be used for estimating the tuning parameters. According to the location of simulated missing values, $D_C^i$ is dived into $D_{training}$, $D_{testing}$. Cross-validation is used to automatically select the values of the tuning parameters yielding the smallest imputation error. Meanwhile, the SMOTE-NC method is applied to imbalanced and categorical features. At last, an ALWRF model (donate as $F_{f_i}$) is built so that the feature $f_i$ is the targeted variable and the rest of the features without missing values are predictive features. If the targeted variable is a numerical variable, the built forest is a regression forest. If the targeted variable is a categorical variable, a classification forest is built. To compute the optimal values of the tuning parameters, the optimization procedure is described in section 3.2.

Step 5. Use the optimal values of the tuning parameters to build an ALWRF model (donate as $F_{f_i}^{optimal}$), and then use it to impute missing values at the $f_i$ feature in $D_I^i$.

Step 6. Repeat steps 3 to 5 until all features with missing values are traversed.

The proposed SncALWRFI method for missing values is shown in algorithm 2.

**Algorithm 2** The proposed imputation method: SncALWRFI

$D$: A data set with missing values $\widetilde{D}$: Data set has been imputed

$M \leftarrow$ Calculate indicator matrix

$D^{'} \leftarrow$ Using mean or mode values as an initial imputation

$\widetilde{F} \leftarrow$ The sorted feature set by missing rate in descending order

**for** $f_i \in \widetilde{F}$ **do**

  $D_C^i, D_I^i \leftarrow$ Divide dataset according to $M$;

  `// optimal parameter`

  **for** *Cross-validation* **do**

    $D_{training}$, $D_{testing} \leftarrow$ Randomly generate missing values in $D_C^i$

    **for** $n$,$m$,$s$,$\gamma$,$knn$,$irt$ **do**

      **if** $f_i$ *is categorical and* $ir > irt$ **then**

        $D_{training} \leftarrow$ Use SMOTE-NC to oversample and update $D_{training}$

      **end**

      $F_{f_i} \leftarrow$ Build ALWRF for the feature $f_i$

      $loss \leftarrow$ Use $D_{testing}$ to compute the loss value of $F_{f_i}$

    **end**

  **end**

  $F_{f_i}^{optimal} \leftarrow$ Using optimal parameters to build the model

  $D_I^i \leftarrow$ Use $F_{f_i}^{optimal}$ to impute and update $D_I^i$

  $\widetilde{D} \leftarrow$ Update $\widetilde{D}$ using $D_I^i$

**end**

## 3.2 Hyperparameter Optimization

The proposed imputation method SncALWRFI requires the tuning parameters to be specified including the parameters in random forest and the parameters in SMOTE-NC. Bayesian optimization(BO) [145] is a state-of-the-art

optimization framework for the global optimization of expensive black-box functions [145, 110], which can find the optimal value through only a small number of samples. Compared with traditional optimization methods, it does not need the explicit expression of the function. Therefore, Bayesian optimization is employed to search best parameters in our work. In our work, the goal is to improve the predictive performance of the proposed model on both classification and regression tasks where the optimization functions are different. In the classification task, the output can be two or more classes. Therefore, a Confusion Matrix with four different combinations of predicted and actual values commonly used to evaluate classifier performance, as shown in Figure 3.2.

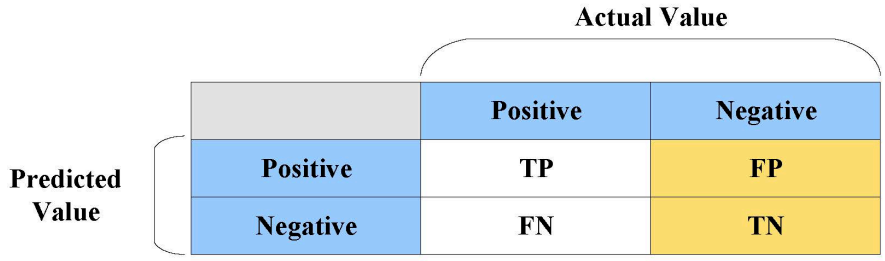



**Figure 3.2:** Confusion Matrix

where TP(True Positive) means that our model predicted positive and it's true; TN(True Negative) means that our model predicted negative and it's true; FP(False Positive) means that our model predicted positive and it's false; FN(False Negative) means that our model predicted negative and it's false. Based on Confusion Matrix, accuracy is employed as the optimization function for classification tasks because it can present how many times our model was correct overall. Accuracy can be computed by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.11}$$

where high accuracy values mean better classification performance. On the other hand, in the regression task, Mean Squared Error (MSE) represents the average squared residual. As the data points fall closer to the regression line, the model has less error, decreasing the MSE. A model with less error produces more precise predictions. The MSE can be calculated by

$$MSE(D) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \tag{3.12}$$

where $f(x_i)$ is the prediction value and $y_i$ is the real value. Due to MSE reflecting the overall deviation of the predicted and true values, the smaller $MSE$ is better. As the low MSE values mean better performance, negative MSE is applied as the optimization function for regression tasks.

### 3.2.1 Bayesian Optimization: ALWRF

Firstly, Bayesian optimization resorts to tuning hyper-parameters for ALWRF. The optimization process is similar to a random forest. The hyperparameters include the number of decision trees in the random forest $n$, the size of the predictor variables subset $m$, minimum sample split $s$, and the interval for updating weights $\gamma$. The default values of hyperparameters are $n = 100$, $m = \sqrt{M}$($M$ is the number of predictor variables), $s = 2$. The ranges of value for hyperparameters are $n \in Range(50, 500, 50)$, $m \in (0.1, 0.999)$, $s \in [2, 25]$

and $\gamma \in \{10, 20, 30, 40\}$, respectively. Here, $m$ is a fraction and it means that $m$ percentage features are considered at each split. Based on the analysis of the 3.1 section, the model prediction accuracy and negative MSE on the test set are chosen as optimization functions. Specifically, the Bayesian optimization process for ALWRF works as follows:

Step 1. Select five sample points randomly in the hyperparameters space and calculate the prediction accuracy or negative MSE of the ALWRF. The five samples are used as the training set;

Step 2. Obtain a new sample point by optimizing the acquisition function and calculating the acquisition function value at the new sample point;

Step 3. Add the new sample point into the training set and update the posterior distribution of the function;

Step 4. Repeat the above steps until reaching the limit of iterations.

In addition, in order to optimize the hyper-parameters of the ALWRF, the dataset is divided into training data, validation data, and testing data. Training data is applied to train the ALWRF model. Validation data is used to tune hyperparameters. The performance of ALWRF is evaluated using testing data. The flow chart of the Bayesian optimization process for the ALWRF is shown in Figure 3.3.

### 3.2.2 Bayesian Optimization: SncALWRFI

Similarly, Bayesian optimization is also employed for hyperparameters of the proposed SncALWRFI method. However, the bayesian optimization process

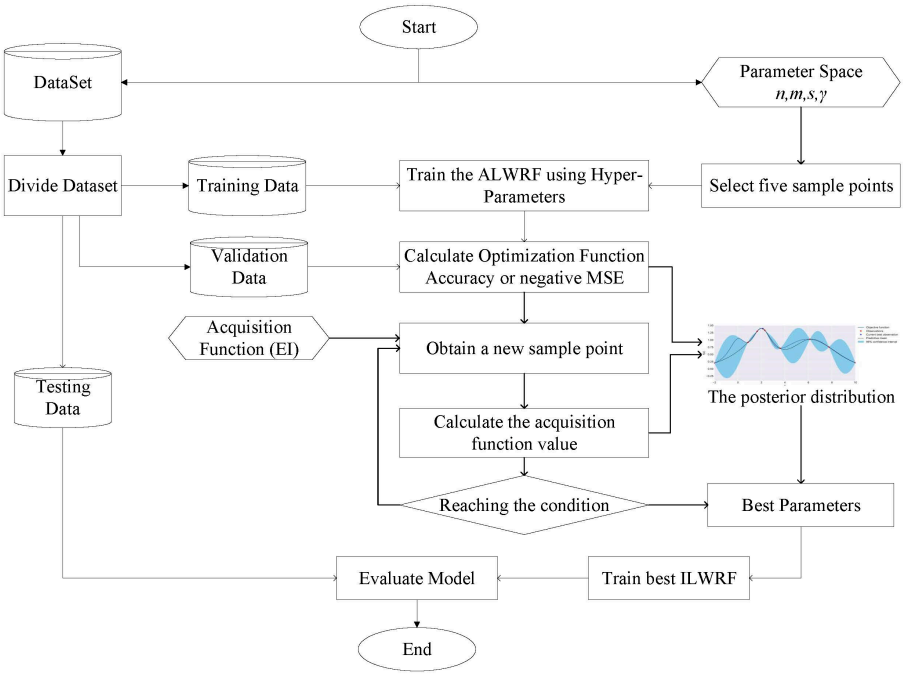



**Figure 3.3:** The bayesian optimization process for the ALWRF.

for the SncALWRFI has three differences from ALWRF. Firstly, as SncALWRFI pays attention to missing values, the optimization function is different and it needs to consider both categorical and numerical features. Therefore, the Proportion of Falsely Imputed Categories ($PFC_{cat}$) is employed as a performance measure for categorical variables, while the Mean of Squared Imputation Errors for numerical values ($MSIE_{num}$) is used as a performance measure for continuous variables.

$$PFC_{cat} = \frac{1}{N} \sum I(x_{ij} \neq x_{ij}^{'}) \tag{3.13}$$

where $I(.)$ is an indicator function, which is 1 when the predicted value and the true value are the same. In addition, the $MSIE_{num}$ can be calculate by

$$MSIE_{num} = \frac{1}{N}(x_{ij} - x_{ij}^{'}) \tag{3.14}$$

where $N$ is the number of numerical missing values, $x_{ij}$ is the true value in the complete data matrix, and $x_{ij}^{'}$ is the corresponding imputed value. Then the optimization function of the SncALWRFI is the sum of $PFC_{cat}$ and $MSIE_{num}$. Secondly, missing values should be randomly introduced in validation data. In detail, we temporarily set as missing some of the available values in the full data matrix, and these missing records make up the validation data for estimating hyperparameters. The third difference is that more parameters should be considered because of the SMOTE-NC method. The additional parameters include $knn$ and $irt$ which are the number of neighbors in SMOTE-NC and the threshold of imbalance rate respectively. The ranges of value for hyperparameters are $knn \in \{3, 5, 20\}$ and $irt \in \{2, 5, 10\}$.

### 3.3 Experiments for Adaptive Laplacian Weight Random Forest

At first, two experiments are conducted to evaluate the performance of the proposed adaptive Laplacian weight random forest. As two category tasks including classification and regression tasks can be applied in the random forest model, we used 4 public medical datasets and 4 public datasets to evaluate the classification and regression performance of the AILWRF method, respec-

tively. In this experiment, feature scaling is not required since the proposed and compared methods are tree-based models. All models were implemented using Python Language and the configuration of the experimental environment is Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz, 8 GB RAM.

### 3.3.1 Classification Task

As we introduced in section 3.1, the classification task is that learn how to assign a class label to samples. Therefore, prediction accuracy and AUC are employed as performance measurements for the classification task. The AUC-ROC curve is a common performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. A higher AUC means that the model has a higher capability to predict class 0 as 0 and class 1 as 1. By analogy, a higher AUC in disease prediction shows the model has a better ability at distinguishing between patients with the disease and no disease.

In the classification task, the information of 4 public medical datasets is shown in Table 3.2. Specifically, three datasets focus on hypertension prediction including Men's dataset, Women's dataset, and the NHANES dataset. Men's dataset and Women's dataset are freely available in a web repository for reproducible purposes [146, 147]. The predictive variables included in these datasets were Body Mass Index (BMI), WC (Waist Circumference), HC (Hip Circumference), and WHR (Waist-to-Height Ratio). NHANES dataset [62] is a subset

of National Health and Nutrition Examination Survey (NHANES) from 2007 to 2017. This dataset can be used to predict the occurrence of hypertension using 7 features that associate with hypertension, such as gender, race, age, smoking, BMI, diabetes, and kidney conditions. The fourth datasets is called Pima dataset [148], which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

| Dataset | Total Samples | Total Variables | Categorical | Numerical |
|---|---|---|---|---|
| Men's dataset | 175 | 7 | 2 | 5 |
| Women's dataset | 224 | 7 | 2 | 5 |
| Pima dataset | 768 | 9 | 1 | 8 |
| NHANES dataset | 24,434 | 8 | 8 | 0 |

**Table 3.2:** The data information for the ALWRF classification experiment

In order to evaluate the performance of the ALWRF after Bayesian optimization (BO-ALWRF), random forest(using default parameters) and random forest after Bayesian optimization (BO-RF) methods are employed for comparison. First, the three methods were performed 20 times, and then four boxplots were used to present their accuracy values across the four datasets, as shown in Figure 3.4.

In Figure 3.4, the distribution of accuracy values for RF, BO-RF, and BO-ALWRF is presented. The results showed that the median of RF was the lowest in all datasets. After hyperparameter optimization, the accuracy values increased, especially in the NHANES dataset. As we expected, BO-ALWRF showed the

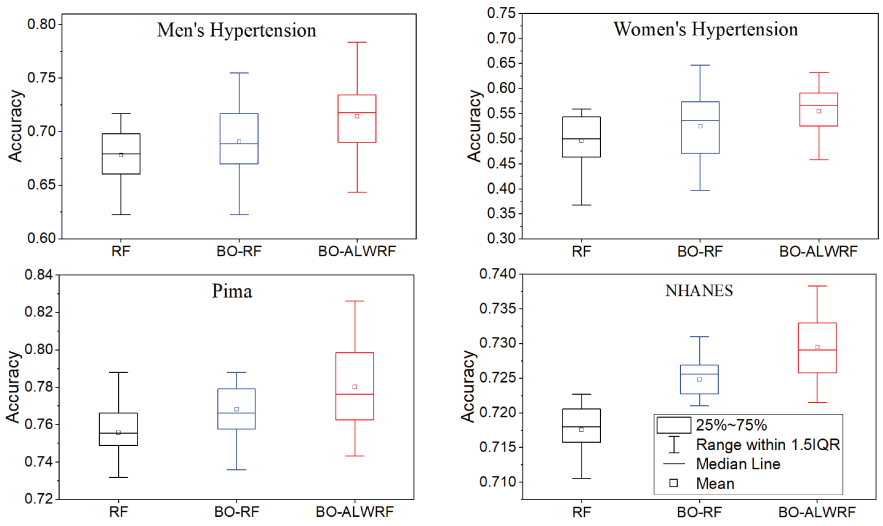



**Figure 3.4:** The distribution of accuracy in the ALWRF classification experiment

best performance on the four datasets, which indicates that introducing adaptive Laplacian weights can improve accuracy. Further, the AUC-ROC curve is also resorted to measuring the separability of three methods, as shown in Figure 3.5.

According to Figure 3.5, RF shows AUC values of 0.485, 0.538, 0.799, and 0.793 on Men's dataset, Women's dataset, Pima dataset and NHANES dataset, respectively before Bayesian optimization. The AUC values of the four datasets were boosted by hyperparameter optimization. In addition, the classification capability of the BO-ALWRF improved (0.606, 0.635, 0.848, and 0.767), based on the results of the RF and BO-RF models. In summary, the proposed adaptive Laplacian-weighted random forest method exhibits better performance in terms of accuracy and AUC metrics for classification tasks in four datasets. On the other hand, since missing values may be continuous features, we also need to pay attention to their performance in regression tasks.

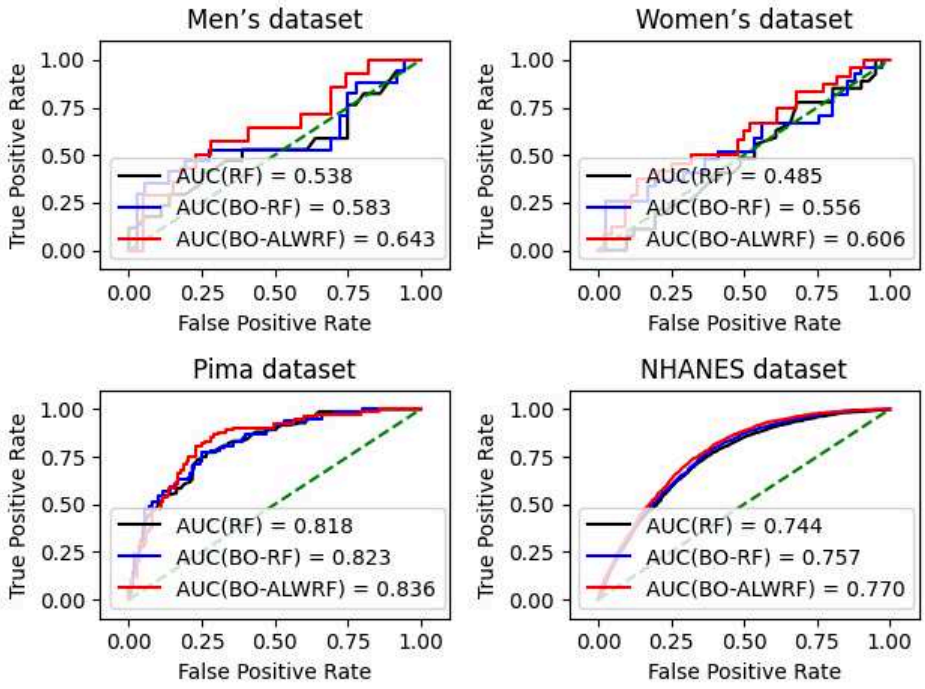



**Figure 3.5:** The AUC-ROC curve in the ALWRF classification experiment

### 3.3.2 Regression Task

Further, four public datasets are used for validating the performance of the proposed ALWRF method in regression tasks. Specifically, insurance cost and dataset include 7 variables in terms of age, sex, BMI, children, smoker, region, and charges, where a charge is the target variable and it represents individual medical costs billed by health insurance. The second dataset is related to life expectancy and it consists of 22 columns and 2938 rows which means 20 predicting variables. Both insurance cost and life expectancy datasets are available on the Kaggle website. The other two datasets are related to red and white variants of the

Portuguese "Vinho Verde" wine [149]. The quality is the target variable and the other 11 variables are predicting variables. The information of these four datasets is shown in Table 3.3.

| Dataset | Total Samples | Total Variables | Categorical | Numerical |
|---|---|---|---|---|
| Insurance cost dataset | 1,338 | 7 | 3 | 2 |
| Life expectancy dataset | 2,838 | 22 | 2 | 20 |
| Red Wine dataset | 1,599 | 12 | 0 | 12 |
| White Wine dataset | 4,898 | 12 | 0 | 12 |

**Table 3.3:** The data information for the ALWRF regression experiment

Similarly, random forest(using default parameters) and random forest after Bayesian optimization (BO-RF) methods are employed for comparison and each method is performed 20 times. In regression tasks, MSE (equation (3.13)) and $R^2$ (coefficient of determination) are used as performance indicators. $r^2$ represents the proportion of variance that has been explained by the independent variables in the model and provides an indication of goodness of fit and therefore a measure of how well-unseen samples are likely to be predicted by the model, through the proportion of explained variance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{3.15}$$

where $f(x_i)$ is the prediction value, $y_i$ is the real value and $\overline{y}$ is the mean of $y$. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). As such variance is dataset dependent, $R^2$ may not be meaningfully compared across different datasets. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant

model that always predicts the expected (average) value of $y$, disregarding the input features, would get a score of 0. Firstly, in order to evaluate the predictive accuracy, we used a bar chart to present the average MSE of three methods over 20 runs, as shown in Figure 3.6. At the same time, we also draw a boxplot to present the distribution of $R^2$ over 20 runs, as shown in Figure 3.7.

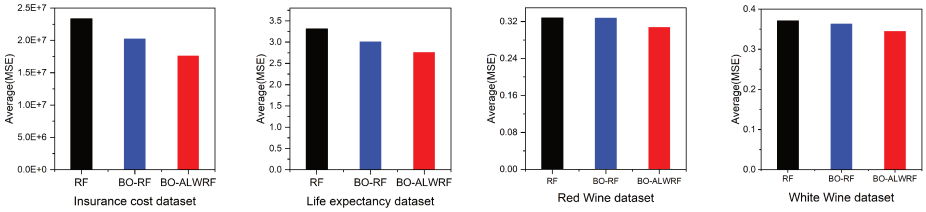



**Figure 3.6:** The average of MSE in the ALWRF regression experiment

According to 3.6, we can observe that RF shows the worst mean MSE across the four datasets. As expected, although the performance of the regression task can also be improved by optimizing the parameters of the random forest, the proposed ALWRF can obtain the lowest average MSE over 20 runs on the four datasets comparing RF and BO-RF. In addition, the ALWRF similarly shows the best medium of $R^2$ across four datasets in 3.7. Although the $R^2$ of BO-RF in the red wine dataset is close to RF, the proposed method can also improve the performance, which indicates that the proposed method is more robust. Overall, the proposed ALWRF method outperforms RF and BO-RF on two common metrics for regression tasks, MSE and R2, suggesting that the use of adaptive Lapland weights can improve the predictive power of random forests.

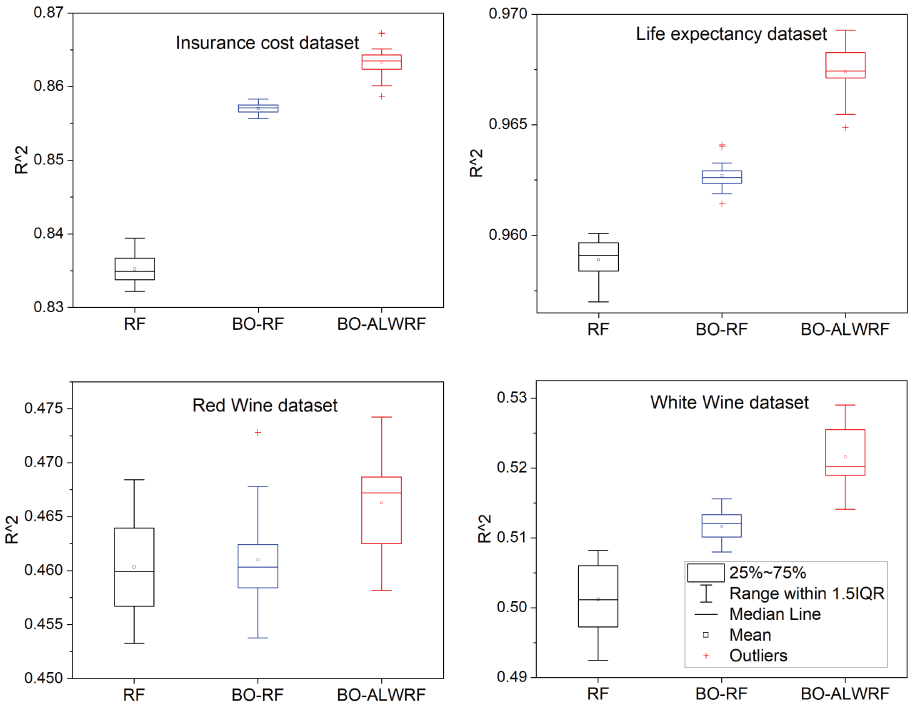



**Figure 3.7:** The distribution of $R^2$ in the ALWRF regression experiment

## 3.4 Experiments for Missing Value Imputation

In this section, we used 16 datasets to evaluate the performance of the proposed SncALWRFI imputation method, where types of experiments are applied in terms of imputation error and imputation effectiveness in classification tasks. In the first category, imputation error is computed by comparing the difference between imputation values and real values. In the second category, we compared the performance using both complete data and imputed data that deal with missing values by a variety of imputation methods.

### 3.4.1 Imputation Error

In order to evaluate the true imputation errors of imputation methods, only complete datasets are used in this experiment. If a dataset has naturally missing values, we discard incomplete rows. Specifically, missing values are then introduced into each data completely randomly at a specific level. and then imputation methods are employed to impute missing values. Finally, imputation accuracy is evaluated by comparing imputed values and real values. The experimental procedure is shown in Figure 3.8.

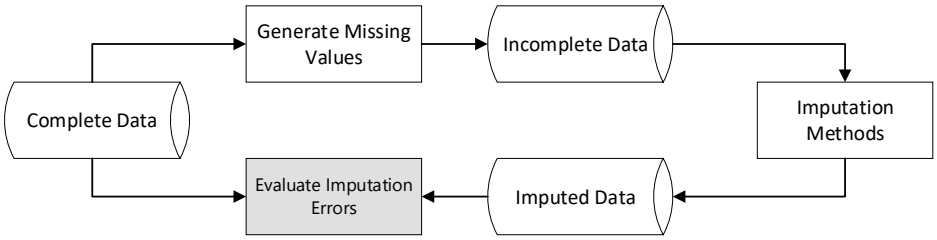



**Figure 3.8:** The experiment flow of simulation missing values for imputation error.

Here, three benchmarks are applied including k-Nearest Neighbors Imputation (kNNI) [109], Random Forest Imputation or MissForest (RFI) [111] and Weighted Nearest Neighbor Imputation using Selected Variables ($wNNSel_{mix}$) [112]. In the kNNI approach, an imputed value is obtained by taking the average of the values of k candidate samples, called neighbors, chosen based on a distance measure. In order to apply for mixed-type data, Gower's distance [150] is employed as the distance measure. The RFI approach is applicable to categorical as well as continuous data even in the case of a high number of predictors. This approach firstly uses simple imputation like mean imputation as an initial method,

and then improves the imputed data by random forest model on each successive iteration. Further, the $wNNSel_{mix}$ approach makes practical and effective use of the information on the association among the variables to improve imputation accuracy. In this experiment, three missing percentages of 10%, 20%, and 30% are simulated. For each missing percentage, we repeat each configuration 200 times to reduce noise from simulating missing values. In order to compare the performance of different imputation procedures, $PFC_{cat}$ and $MSIE_{num}$ are used as performance measures for categorical and continuous variables, respectively.

Specifically, five public datasets are used including German Breast Cancer Study Group 2 (GBSG2) data, Hepatitis dataset (Hepatitis), Body Mass Index dataset (BMI), Cars dataset(Cars), and Automobile dataset (Automobile). The Hepatitis dataset is from UCI Machine Learning Repository [151] and the other four datasets can be found in the R package. In order to compare performance with $wNNSel_{mix}$ (denoted as $wNN$), the same experimental datasets [112] are used and their information is shown in Table 3.4. In addition, experimental datasets are normalized by StandardScaler in this experiment because kNNI and $wNNSel_{mix}$ methods are easily affected by data scalar. Here, IRs show the range of imbalance rates for features.

| Dataset | Total Samples | Used Samples | Total Variables | Categorical | Numerical | IRs |
|---|---|---|---|---|---|---|
| GBSG2 | 686 | 100 | 10 | 3 | 7 | [1.29,5.48] |
| Hepatitis | 155 | 155 | 19 | 13 | 6 | [1.03,8.69] |
| BMI | 152 | 152 | 6 | 2 | 4 | [1.14,2.71] |
| Cars | 93 | 82 | 24 | 6 | 18 | [1.05,12.6] |
| Automobile | 205 | 155 | 24 | 9 | 15 | [1.0,64.0] |

**Table 3.4:** The information of five public datasets.

According to Table 3.4, it is easy to observe that some of the existing features in the four datasets are unbalanced, especially in the Automobile dataset. The experiment results are shown in Table 3.5.

| Dataset | MR | Total MSIE | | | | Categorical PFC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *kNNI* | *RFI* | *wNN* | **Proposed** | *kNNI* | *RFI* | *wNN* | **Proposed** |
| GBSG2 | 10% | 1.4156 | 1.0012 | 0.8524 | **0.8141** | 0.3540 | 0.2820 | 0.2140 | **0.1937** |
| | 20% | 1.5560 | 1.0700 | 0.9492 | **0.8713** | 0.4075 | 0.3005 | 0.2540 | **0.2247** |
| | 30% | 1.5257 | 1.0753 | 0.9411 | **0.9104** | 0.4097 | 0.3040 | 0.2490 | **0.2301** |
| Hepatitis | 10% | 1.3805 | 1.3098 | 1.0108 | **0.917** | 0.3912 | 0.3600 | 0.3169 | **0.2723** |
| | 20% | 1.3782 | 1.3425 | 1.1111 | **1.0312** | 0.3962 | 0.3653 | 0.3259 | **0.2914** |
| | 30% | 1.4429 | 1.3622 | 1.2050 | **1.112** | 0.4060 | 0.3690 | 0.3467 | **0.3009** |
| BMI | 10% | 0.9780 | 0.8623 | 0.7728 | **0.7646** | 0.2928 | 0.3469 | 0.2543 | **0.2481** |
| | 20% | 1.1857 | 1.0588 | 1.0740 | **0.968** | 0.3902 | 0.4264 | 0.3475 | **0.3348** |
| | 30% | 1.3215 | 1.1857 | 1.1731 | **1.1118** | 0.3904 | 0.4367 | **0.3575** | 0.3614 |
| Cars | 10% | 0.4860 | 0.2854 | 0.1735 | **0.1677** | 0.2125 | 0.1450 | 0.1250 | **0.1167** |
| | 20% | 0.5462 | 0.3038 | 0.2058 | **0.2023** | 0.2275 | 0.1550 | 0.1403 | **0.1368** |
| | 30% | 0.6335 | 0.3821 | 0.2368 | **0.2340** | 0.2308 | 0.1562 | **0.1430** | 0.1456 |
| Automobile | 10% | 0.4412 | 0.1756 | 0.1824 | **0.1579** | 0.2537 | 0.0981 | 0.0881 | **0.0745** |
| | 20% | 0.4637 | 0.1872 | 0.1910 | **0.1732** | 0.2819 | 0.1081 | 0.0942 | **0.879** |
| | 30% | 0.4727 | 0.1978 | 0.2060 | **0.1868** | 0.2978 | 0.1185 | 0.1050 | **0.941** |

**Table 3.5:** The experiment results of imputation errors.

In Table 3.5, the total error is listed on the left and the error for categorical variables is on the right. The error for numerical variables can be calculated using the total error minus the error for nominal variables. From Table 3.5, the imputation quality is affected by the percentage of missing values. Especially in the BMI and Cars datasets, imputation accuracies of all approaches deteriorated rapidly

with increasing missing values. While compared with other methods, our proposed SncALWRFI method has the best total error regardless of the percentage of missing data. The KNNI method always provides poor imputation because the other three methods seem to use the correlation among covariates for imputation to provide better imputation results. However, the imputation performance of the proposed method for categorical variables is lower than that of the $wNNSel_{mix}$ method in the BMI and Cars datasets at the 30% missing rate. The main reason is that the features in the BMI dataset are nearly balanced, and only a few samples are available to build the random forest model for the Cars dataset. As expected, our proposed SncALWRFI method outperforms the other three models overall, which is attributed to adaptive Lapland weights and oversampling techniques.

### 3.4.2 Imputation Effectiveness in Classification Tasks

The imputation error describes how accurately the imputation of missing values is done by the imputation techniques. However, it does not guarantee that a good imputation always improves data quality for a data mining task such as classification [152]. Therefore, the main objective of this section is to evaluate the effectiveness of the imputation techniques for data mining by applying several classifiers on the original data set, imputed data set and the data sets have missing values. As the prediction accuracy of a classifier can be used to evaluate the impact of the imputation of missing values [152], an evaluation model is built in order to find the prediction accuracy as the effectiveness of an imputation technique. In addition, since the true value of missing data is unknown in the

real world, imputation effectiveness in classification tasks is more important than the evaluation of imputation error. Therefore, we paid more attention to this experiment and used two types of dataset terms complete data and incomplete data. different missing rates can be easily simulated in complete data, while it can not replace real missing values. The overall block diagram of the experiment flow is shown in Figure 3.9.

In this experiment, a dataset is firstly divided into two sub data sets namely a testing data set and a training data set. As we used two types of experiment data including complete data and incomplete data, we then need to introduce missing values in the complete training dataset. Next, deletion and imputation techniques are employed to deal with missing values in both the training dataset and the testing dataset. Further, some popular classifiers are applied to each complete data set and thereby build prediction models. Finally, for each prediction model, we calculate its prediction accuracy by applying it to the testing data set.

Specifically, five missing rates are adopted including 10%, 15%, 20%, 25%, and 30% in this experiment. For each missing rate, each configuration is also repeated 200 times. In addition, in order to compare our model with more models that are suitable for mixed-type data, we implemented four imputation methods using a similar strategy with [111] based on the CART tree [136], the Adaptive Boosting Decision Tree (ABDT) [126], the Gradient Boosting Decision Tree (GBDT) [153] and Multi-Layer Perceptron (MLP) [154]. Therefore, seven imputation methods including k-nearest neighbors imputation (kNNI), the random forest imputation (RFI), the decision tree imputation (DTI), the AdaBoost decision

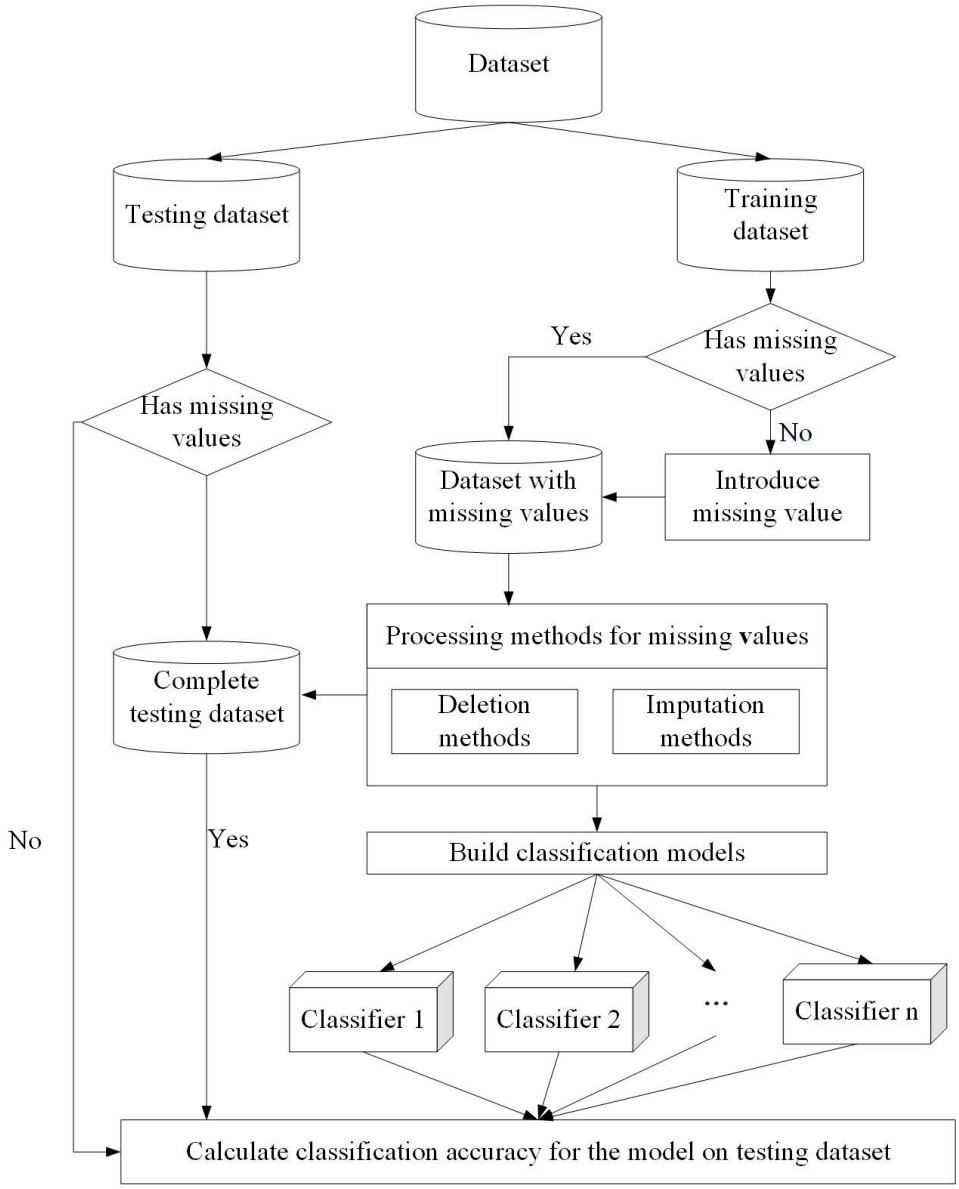



**Figure 3.9:** The overall block diagram of the imputation effectiveness experiment flow

tree imputation (ABDTI), the gradient boosting decision tree imputation (GBDTI), the multi-layer perceptron imputation (MLPI) and the proposed method SncALWRFI are used in this experiment. Moreover, three classifiers with differ-

ent structures are adopted as evaluation classifications, namely Linear Regression (LR) [155], Naive Bayes Network (NB) [156] and Support Vector Machine (SVM) [157].

Firstly, we used six complete medical datasets to evaluate imputation effectiveness including Statlog heart data (Statlog), heart failure by cardiovascular diseases (Heart Failure), early-stage diabetes risk prediction dataset (Diabetes Risk), contraceptive method choice dataset (CMC), the dataset for estimating obesity levels based on eating habits and physical condition (Obesity) [158], and cardiovascular disease dataset (Cardiovascular). The Obesity and Cardiovascular datasets are from the Kaggle platform and the other four datasets are from UCI Machine Learning Repository [151]. The information of these six experiment datasets is shown in Table 3.6.

| **Dataset** | **Total Samples** | **Used Samples** | **Total Variables** | **Categorical** | **Continuous** | **IRs** |
|---|---|---|---|---|---|---|
| Statlog | 270 | 270 | 14 | 7 | 7 | [2.1, 68.5] |
| Heart Failure | 299 | 299 | 13 | 6 | 7 | [1.32, 2.11] |
| Diabetes Risk | 520 | 520 | 17 | 16 | 1 | [1.02, 4.91] |
| CMC | 1,473 | 1,473 | 10 | 8 | 2 | [1.89,20.43] |
| Obesity | 2,110 | 2,110 | 17 | 9 | 8 | [1.02, 225.71] |
| Cardiovascular | 70,000 | 5,000 | 12 | 7 | 5 | [1.0,17.6] |

**Table 3.6:** The information of six public medical datasets.

In order to compare the overall performance of imputation methods, we calculated the average accuracy of five levels of missing rate in three classifiers for each dataset, as shown in Figure 3.10.

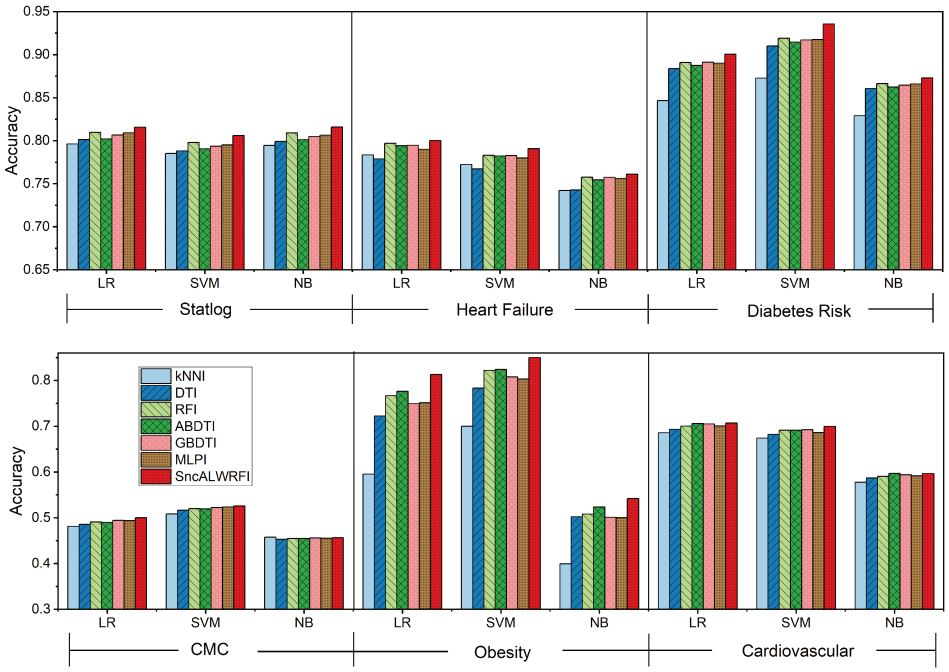



**Figure 3.10:** The experiment results of for the total missing rates.

According to Figure 3.10, we observed that the classification results of the NB classifier are lower than LR and SVM classifiers in most situations because it assumes that each feature makes an independent and equal contribution to the target class. Additionally, the kNN imputation based on Gower's distance performs the worst, followed by a single decision tree. The reason is that the kNN imputation ignores the correlation between covariates, while the single decision tree imputation method is susceptible to noise. By contrast, the prediction accuracy of ensemble models is relatively stable and similar. Although the MLPI performs well, it has high complexity. Overall, our proposed method outperforms other methods in all datasets. Especially in the Obesity dataset with more imbalanced categorical features, the proposed method has obvious advantages, as a

result of the advanced oversampling algorithm. Compared with RFI, our proposed method always performs better, which indicates that the prediction model based on the proposed adaptive Lapland weights can improve the quality of the data. Next, to analyze the performance of our proposed method on different missing rates, we calculated the average performance in three classifiers for imputation methods, as shown in Figure 3.11.

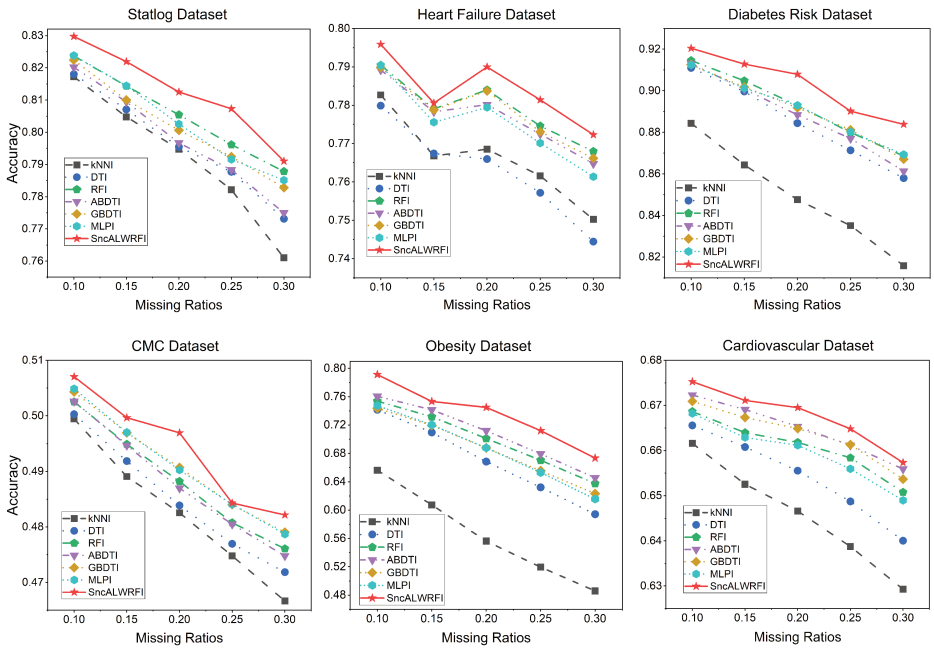



**Figure 3.11:** The experiment results of three classifiers vales.

In Figure 3.11, the prediction accuracy of imputation methods drops sharply as the missing rate increases. Additionally, we notice that the kNN imputation performed poorly on the Diabetes Risk and Obesity datasets because these two datasets have more features, and it ignored the correlation of features. As expected, the performance of a single decision tree is always lower than ensemble

models in all missing rates. While our proposed method performs best under different missing rates across all datasets, it shows that our proposed imputation method is robust in different missing rates.

Further, to verify the performance of our proposed imputation method in datasets with true missing values, six public medical datasets from UCI [151] with real missing values are used, including the Cleveland heart disease dataset (Cleveland), Hepatitis, primary tumor dataset (Primary Tumor), chronic kidney disease dataset (Chronic Kidney), Thyroid dataset and Framingham heart study cohort dataset (Framingham). Their information is shown in Table 3.7.

| **Dataset** | **Total Samples** | **Total Variables** | **Categorical** | **Continuous** | **IRs** | **% MV** |
|---|---|---|---|---|---|---|
| Hepatitis | 155 | 19 | 13 | 6 | [1.03,8.69] | 5.39 |
| Cleveland | 303 | 14 | 9 | 5 | [2.06,37.75] | 0.14 |
| Primary Tumor | 339 | 18 | 17 | 1 | [1.10,47.43] | 3.69 |
| Chronic Kidney | 400 | 25 | 23 | 2 | [1.91,22.31] | 10.09 |
| Thyroid Dataset | 2,800 | 30 | 8 | 22 | [1.0,199.0] | 5.42 |
| Framingham | 4,238 | 16 | 7 | 9 | [1.33,168.52] | 0.95 |

**Table 3.7:** The information of datasets with real missing values.

In order to show the distribution of missing values in the datasets, we use the missing matrix to identify where missing values occur in real cases. The missing matrix of these six datasets is shown from Figure 3.12 to Figure 3.17.

When data is present, the plot is shaded in black, and when it is absent the plot is displayed in white.

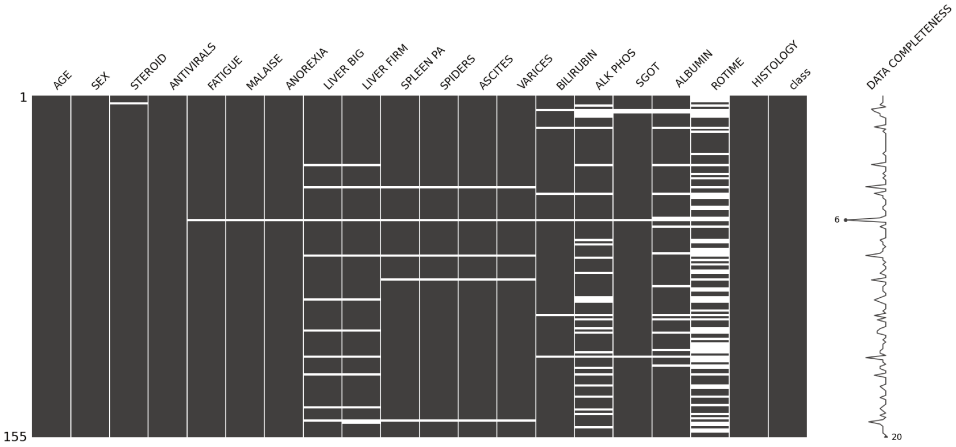



**Figure 3.12:** The missing matrix of Hepatitis Dataset

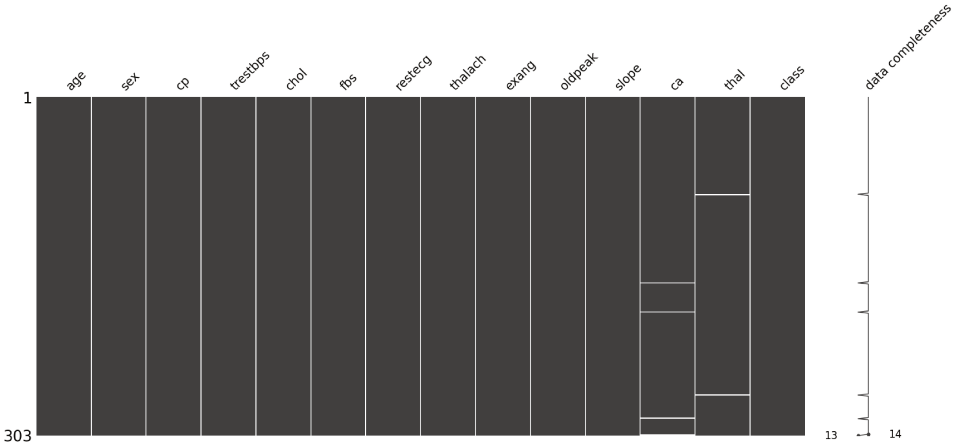



**Figure 3.13:** The missing matrix of Cleveland Dataset

As seen in the plot, the Chronic Kidney dataset shows the largest total missing rate (10.09%) and the Cleveland dataset shows the smallest total missing rate (0.14%). In addition, the missing values are widely distributed in the Hepati-

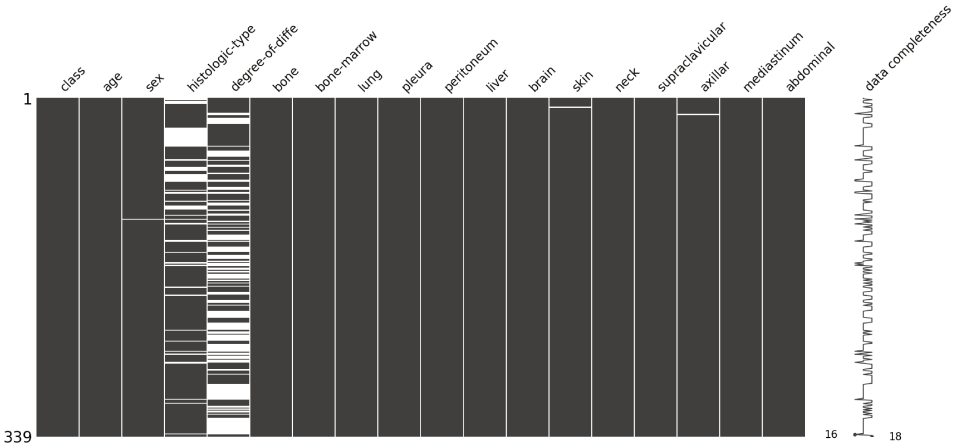



**Figure 3.14:** The missing matrix of Primary Tumor Dataset

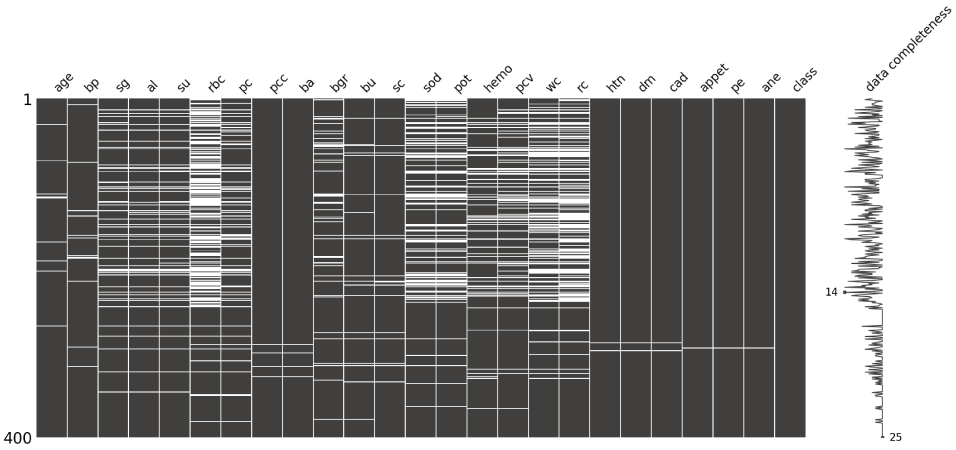



**Figure 3.15:** The missing matrix of Chronic Kidney Dataset

tis dataset and Chronic Kidney dataset. For the other four datasets, the missing values are concentrated in some columns. Specifically, We also adopted the same six imputation methods as benchmarks, while adding two popular deletion methods (LD and PD) for comparison. As deletion methods are limited in some scenes, we computed instance-missing rates and column-missing rates of these datasets. The instance missing rate is the percentage of instances in the data set which

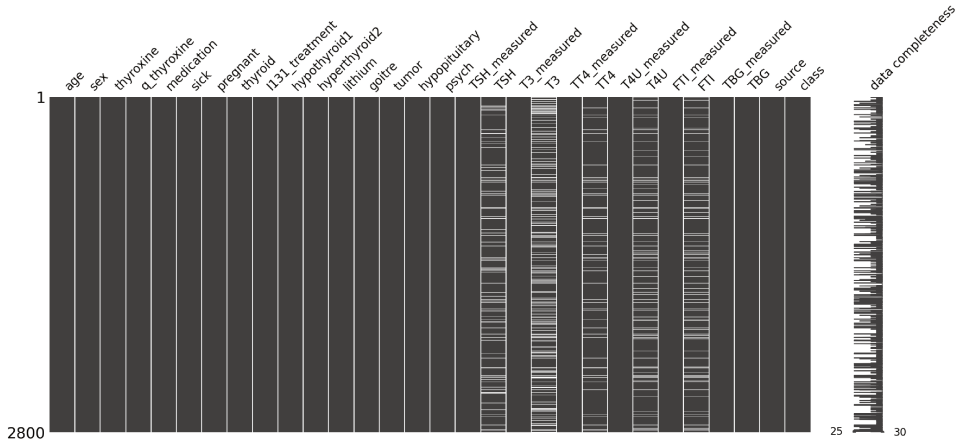



**Figure 3.16:** The missing matrix of Thyroid Dataset

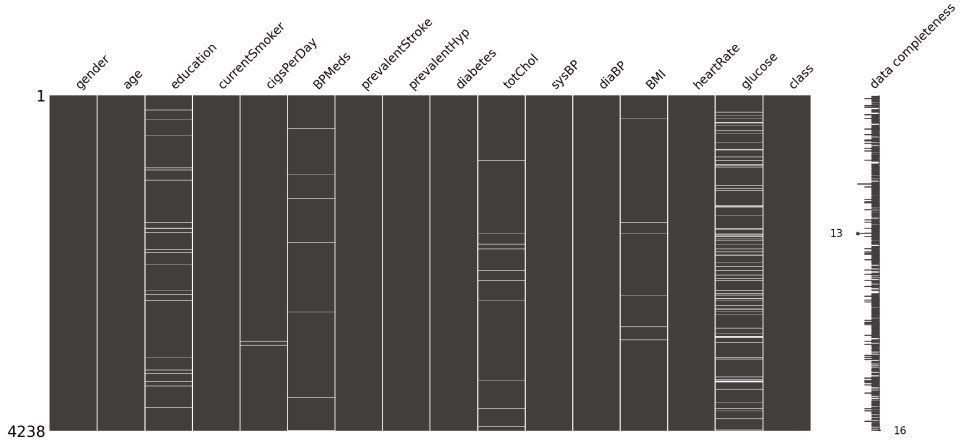



**Figure 3.17:** The missing matrix of Framingham Dataset

have at least one missing value and the column missing rate is the proportion of columns in the dataset that have at least one missing value. The instance missing rates of the Hepatitis dataset, Cleveland dataset, Primary Tumor dataset, Chronic Kidney dataset, Thyroid Dataset and Framingham dataset are 48.39%, 1.98%, 61.06%, 60.5%, 100%, 13.72% respectively and their column missing rates are 75.0%, 14.29%, 27.78%, 96.0%, 16.67%, 37.5% respectively. Therefore, as the

instance missing rates of the Thyroid Dataset is 100%, the LD method can not be used. At the same time, as there are no complete covariates in the Chronic Kidney dataset, the PD method can not be used. Additionally, since LD and PD methods cannot be used directly when missing values occur in test data, we use mean and mode to fill them to ensure using the same size of testing data for all methods. The experiment results are shown in Figure 3.18.

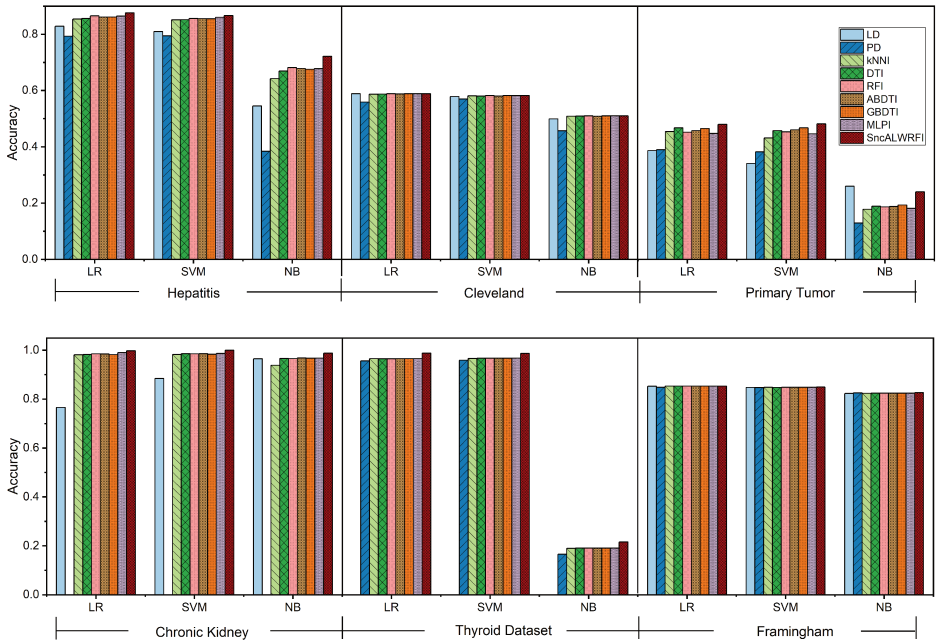



**Figure 3.18:** The experiment results of datasets with real missing values.

Firstly, the performance of all imputation methods and the LD method is close in the Cleveland dataset, which is attributed to the highly lower instance missing rate (less than 1.98%). While the PD method performed badly because it deleted two columns. In this case, the LD seems like a favorable choice. On the other hand, all methods performed similar prediction accuracy in the Framingham

dataset, even for two deletion methods. The reasons are the missing rate is low and missing features play less impact on the predicted outcome. In this case, PD may be a suitable choice. However, for the other four datasets with high instance missing rates, our proposed method consistently shows the best performance, except for the performance of the LD method for the NB classifier on the Primary Tumor Dataset. Although the LD method outperforms our proposed method on the Primary Tumor dataset for the NB classifiers, its performance is highly weaker for LR and SVM classifiers. Generally, when missing values are concentrated in a few features, the PD can be used, but if these features need to be preserved, our proposed method has great competitiveness. To sum up, although our proposed imputation method performs similarly to other methods in datasets with low missing rates, it shows the best imputation effectiveness in datasets with high missing rates. But where the missing rate of the dataset is low, the prediction accuracy of our method is not always better than the deletion method in classification tasks. Therefore, our method is not necessarily the best choice for studies that have low missing rates and focus only on classification accuracy. But for studies with a high missing rate or need to retain more samples, our method can significantly improve data quality.

## 3.5 Summary

As features uniformly and randomly are selected to form a feature subspace in a random forest, features with high quality are not fully utilized. We proposed an improved random forest model, called adaptive Laplacian weight ran-

dom forest (ALWRF), in which features' weights adaptively adjust when building a random forest. Meanwhile, cross-validation and Bayesian optimization are employed to search hyper-parameters. Then eight pubic datasets are used to verify the prediction ability of the ALWRF on the classification and regression tasks. The experiment results show that the ALWRF outperforms random forest and Bayesian optimized random forest.

Missing values is an inevitable problem when mining useful information from medical data. In order to improve the quality of incomplete medical data with the characteristics of imbalance and mixed type, an imputation method (SncALWRF) is proposed based on the ALWRF and the oversampling technology SMOTE-NC. In the experiment for missing values, we first compared the imputation errors of the proposed method with three advanced imputation methods using five small complete data subsets. Experiment results show that the proposed method provides excellent imputation estimates for missing values in categorical and numerical variables.

We then focus on the imputation effectiveness of the proposed imputation method in the classification tasks. We first used six complete datasets with the characteristics of imbalance and mixed type to evaluate the prediction accuracy of the proposed imputation model at different missing rates. Experiment results show although with the increasing of missing ratio, the imputation performance for all imputation methods deteriorates, the decrease is more gradual for the proposed method. At the same time, our method outperforms other imputation methods in the same missing values. We then adopted six public medical

datasets with real missing values and compare them to evaluate the effectiveness of our proposed method in classification tasks and compared them with other 6 imputation methods and 2 deletion methods. Experiment results show when datasets with low missing rates (5%), our model can not always perform well than deletion methods, but it outperforms other imputation methods in the real case study. Therefore, our imputation method can significantly improve data quality for studies with high missing rates or the need to retain more samples.

# Chapter 4. A Stacking-Based Ensemble Approach for Noise Data

## 4.1 Methodology of the Proposed Stacking-Based Approach

It is notoriously difficult to utilize an individual model due to its unidirectionality, domain unity, and inherent quality. In addition, it is challenging to employ a single model to generate more accurate forecasts and attain higher levels of performance due to the noise from attributes and classes. In machine learning, an ensemble is a sort of model that is built by merging the predictions of various individual models. Typically, ensembles increase performance by reducing the mistakes created by each individual model that contributes to the ensemble. Generally, there are two challenges in the ensemble framework in terms of model selection and model fusion.

### 4.1.1 Model Selection

There are many types of research devoted to the selection of meta-leaners. The paper [159] adopted prediction accuracy as an objective, optimized by an artificial bee colony algorithm to collect meta-learners. In [160], the ant colony algorithm was applied to optimize local information, which represented the precisions of the meta-level classifiers to configure stacking ensembles. But the single-

objective optimization algorithms usually adopt a greedy search strategy that easily leads to a local minimum. It doesn't take much accuracy improvement but excess meta-learners. The paper [161] adopted a multi-objective optimization algorithm named non-dominated sorting genetic algorithms-II (NSGA-II) to evolve an ensemble and the result is averaged by each individual. It maximizes the generalization capacity of the ensemble and minimizes its structural complexity simultaneously to get a better ensemble. While the papers [162] and [163] describe that the ideal ensemble is constructed using learners of small error and good diversity. However, rich diversity may cause the predicted value of meta-learners to deviate from the true values, and the improvement of individual accuracy often reduces the diversity of meta-learners, that is, accuracy and diversity are usually conflicting with each other. Further, the selection of meta-learners in the paper [164] followed the NSGA II algorithm to balance the two conflicting objectives in terms of accuracy and diversity. As the NSGA II algorithm randomly initializes the population, optimal individuals are changeable and it requires more meta-learners when generating the offspring in the NSGA II algorithm.

To sum up, accuracy and diversity are two crucial factors that decide the success of stacking. In order to maximize the diversity and the accuracy of ensemble models simultaneously, we proposed a Multi-objective Iterative Model Selection (MoItMS) algorithm. Specifically, accuracy measures the difference between the predicted values and actual values while diversity measures the differences between meta-learners. Suppose there are k individual models which are selected by MoItMS, for the cost function $C_{m_i}$ is defined as:

$$C_{m_i} = E_{m_i} + \lambda D_{m_i} \tag{4.1}$$

where $E_{m_i}$ represents the accuracy, and $D_{m_i}$ represents the diversity. $\lambda$ is the weight and $\lambda = 1$. Here $E_{m_i}$ can be computed by:

$$E_{m_i} = \frac{1}{N} \sum_{j=1}^{N} (p_{m_i}^j - y^j) \tag{4.2}$$

where $y^j$ is the actual values of the $j$-th training sample, and the $p_{m_i}^j$ is the predicted values obtained by the $i$-th meta-learner for the $j$-th training sample. Here, the predicted probabilities are applied to the predicted values instead of class labels. $N$ is the number of samples. According to the paper [164], the correlation $D_{m_i}$ is defined as:

$$D_{m_i} = \frac{1}{N} \sum_{j=1}^{N} ((p_{m_i}^j - p_{avg}^j) \sum_{l \neq i}^{k} p_{m_l}^j - p_{avg}^j) \tag{4.3}$$

where $p_{m_i}^j$ and $p_{m_l}^j$ represent the predicted values of the $i$-th and $l$-th meta-learners for the $j$-th training instance, respectively. $p_{avg}^j$ is the average predicted value of the models in the ensemble. Reference [163] proves that good diversity can be achieved (if there is no bias) when the individual models are negatively correlated, which means the lower the $D_{m_i}$ is, the larger the diversity is. Further, since the two objective functions have different magnitudes, normalization is required so that the algorithm does not favor a larger magnitude. Therefore, the objective functions $f$, $f \in \{E, D\}$ can be normalize by

$$\widetilde{f}_{(X)} = \frac{f_{(X)}}{Z_f} \tag{4.4}$$

where $X$ represents candidate models. $Z_f$ is the normalization factor of each objective function which is the maximum function value in the candidate models. Therefore, the cost function for the ensemble can be the average of these individual model's costs:

$$C = \frac{1}{k}\sum_{i}^{k}(\widetilde{E}_{m_i} + \widetilde{D}_{m_i}) \tag{4.5}$$

where $\widetilde{E}_{m_i}$, $\widetilde{D}_{m_i}$ are the accuracy and diversity of $i$-th meta-learner after normalization, respectively. The small value of $C$ means that the ensemble model combines meta-models with high accuracy and diversity. In order to maximize the accuracy and diversity of the ensemble model, an iterative process is employed to search for the best cost. In detail, the proposed MoItMS algorithm mainly includes six steps:

(1)Firstly, five-fold cross-validation is used to generate a predicted set of a dataset $X$, which will be applied to assess the accuracy and diversity of each individual model.

(2) All candidate models $M = \{m_1, m_2, ..., m_s\}$ is an ensemble model, and then the cost function of each candidate model in this ensemble model is calculated according to equations (1)-(4). The model $\overline{m}_1$ with smallest cost function is selected and add into $\overline{M} = \overline{m_1}$ and it is removed from the candidate models $M = m_1, m_2, ..., m_{s-1}$.

(3) A model $m_i$ is iteratively selected from the candidate models $M$ and then a new ensemble model $G$ is formed combining $m_i$ and $\overline{M}$ The two objective functions values $E_{m_i}$),$D_{m_i}$ of the model $m_i$ in the ensemble model $G$ need to be calculated.

(4) The objective function values of all candidate models are normalized by equation (4). The cost function values of ensemble models are computed by equation (5), and then the model $m_j$ with the smallest cost function is selected and added into $\overline{M}$.

(5) The selected models $\overline{M}$are stacked, and the performance of the ensemble model is evaluated using five-fold cross-validation.

(6) The performance of previously selected models and newly selected models are compared. If the performance is improved and the difference is greater than a threshold $\beta$, then repeat steps 3 to 5. Otherwise, the last added model is pushed out.

In the proposed MoItMS approach, the threshold $\beta$=0.01 is implemented to balance accuracy and complexity, which means that when the performance increase is insufficient, we sacrifice performance and maintain complexity low. In general, the proposed algorithm has lower computational complexity. The computational complexity of the MoItMS is $[s+(s-1)+(s-k)]*O(mn)$, which is lower than the paper [164], $[\sigma*k*G]*O(mn)$. Here, $m$ represents the number of training samples, $n$ is the number of features, $s$ is the number of candidate models, $k$ represents the number of selected models, $\sigma$ is the number of offspring

and $G$ is the generation number. Specifically, the proposed MoItMS is shown in Algorithm 3.

**Algorithm 3** Multi-objective Iterative Model Selection (MoItMS) algorithm

**Input:** Data set $D = (X, Y)$, Candidate models $M = \{m_1, m_2, ..., m_s\}$, the threshold $\beta$, the weight $\lambda$

**Output:** Selected models $\overline{M}$

Selected models $\overline{M} = \{\overline{m}_1, \overline{m}_2, ..., \overline{m}_k\}$ for the ensemble model

Selected models $\overline{M} \leftarrow \emptyset$; The improved performance $p \leftarrow 0$

Using 5 cross-validations to train each candidate model

$Y' = \{Y'_1, Y'_2, ..., Y'_s\} \leftarrow$ The predicted probabilities of all candidate models in validation datasets

$E \leftarrow$ Calculating error of candidate models based on $Y'$ and $Y$ by equation (2)

$\overline{m}_{best} \leftarrow$ The model with the smallest error value

$p \leftarrow$ The performance of $\overline{m}_{best}$ on 5 cross-validation

$\overline{M} \leftarrow \overline{M} \cup \overline{m}_{best}$

**while** $p > \beta$ **do**

  accuracy $E \leftarrow \emptyset$, diversity $D \leftarrow \emptyset$

  **for** *each* $m_i$ *not in* $\overline{M}$ **do**

    $E_i, D_i \leftarrow$ Computing error and diversity of $m_i$ when $m_i$ and $\overline{M}$ form an ensemble model

    $E \leftarrow E \cup E_i$, $D \leftarrow D \cup D_i$

  **end**

  $E', D' \leftarrow$ Normalized $E, D$ by equation (5)

  $C \leftarrow$ Calculating cost values using $E', D'$ by equation (1)

  $\overline{m}_{best} \leftarrow$ The model with the smallest cost value

  $\overline{M} \leftarrow \overline{M} \cup \overline{m}_{best}$, $S_{\overline{M}} \leftarrow$ stacking $\overline{M}$

  $p' \leftarrow$ Evaluating the ensemble model $S_{\overline{M}}$ using 5 cross-validation

  $p \leftarrow (p' - p)$

**end**

$\overline{M}$ remove the last model

### 4.1.2 Model Fusion

In general, ensemble models can be categorized into the homogeneous ensemble and heterogeneous ensembles according to the structure of the component model. Homogeneous ensemble mainly ensemble decision trees in terms of bagging and boosting technologies. Bagging technology [165] often considers homogeneous learners, learns them independently from each other in parallel, and combines them following some kind of deterministic averaging process. Random Forest [166] is the representative model in bagging technology. While boosting technology [167] learns learners sequentially in an adaptative way (a model depends on the previous ones) and combines them following a deterministic strategy, such as Adaptive Boosting (AdaBoost) [168], Extreme gradient boosting (XGBoost) [169] and Light gradient boosting machine (LightGBM) [170]. Further, stacking technology [171] generally considers heterogeneous learners, learns them in parallel, and combines them by training a meta-model to output a prediction based on the different model predictions. Even though different models may have similar error rates, stacking ensembles tend to make different mistakes, since they get different professions. In order to search best leaners for staking ensemble, ACO (Ant Colony Optimization) [172], GA (Genetic Algorithms) [173] and NSGA II (non-dominated sorting genetic algorithms-II) [164] have been resorted. The categories of ensemble models and their representative models are shown in Figure 4.1.

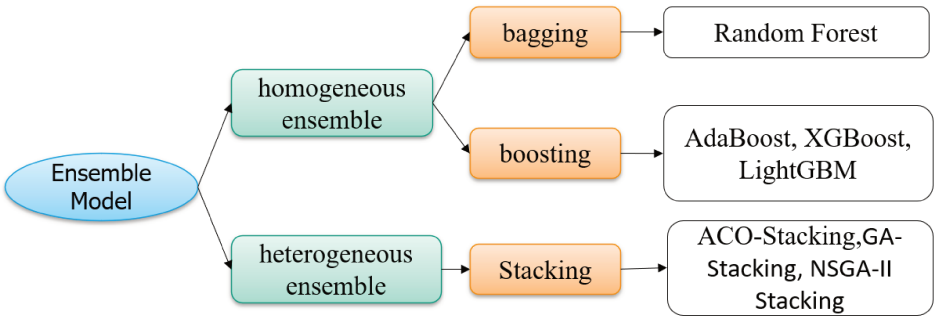



**Figure 4.1:** The categories of ensemble models and representative models

The benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions that have better performance than any single model in the ensemble. Therefore, in order to achieve the best risk prediction, a staking ensemble is employed in our work. The framework of our stacking ensemble approach for hypertension risk estimation of systems under multi-operating conditions is introduced in this chapter, which is illustrated in Figure 4.2. It has three major steps: (i) Model selecting. (ii) Model fusing. (iii) Risk estimating. Firstly, the proposed MoItMS algorithm is applied to select the most suitable meta-learners from candidate models. secondly, the ensemble model is stacked based on these diverse meta-learners. Finally, an extensive analysis was performed to identify the best meta-learner among the employed meta-learners which improves the final prediction accuracy of the proposed system. Therefore, in our stacking approach a neural network with a hidden layer is determined because the neural network model can introduce nonlinearity and one hidden layer can reduce time consumption. Specifically, our

proposed ensemble model includes a two-level classification structure in terms of the base-learner level (level-0 models) and the meta-learner level (level-1 model).

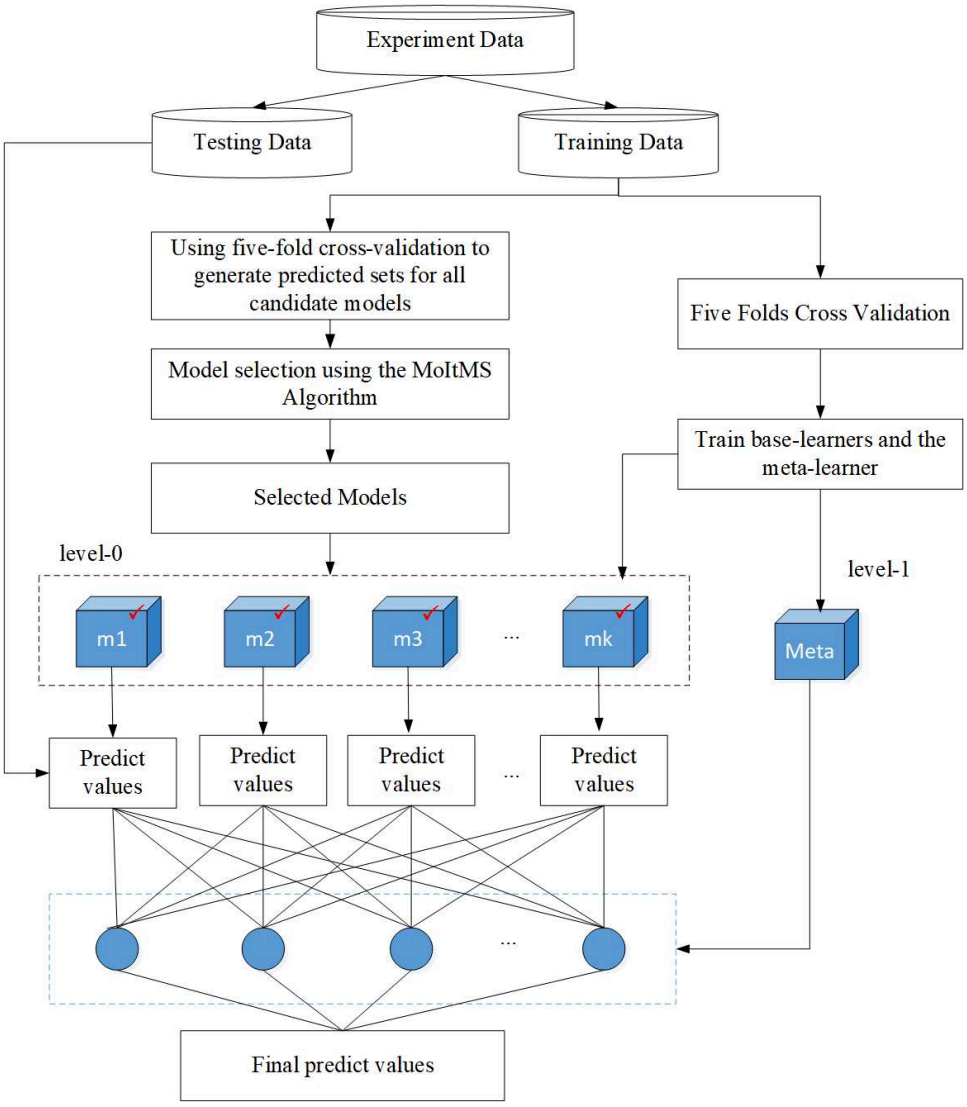



**Figure 4.2:** The framework of the proposed ensemble model

## 4.2 Ensemble Approach Evaluation on A National Health Dataset

### 4.2.1 Dataset Introduction

In order to provide complete access, we used the National Health and Nutrition Examination Survey (NHANES) datasets that were generated and published by the Centers for Disease Control and Prevention (CDC). The dataset includes information on human population statistics (i.e., age and gender), as well as data from examination (i.e., blood pressure and body measures), and questionnaires in terms of disease condition and healthy habits. From 2007 to 2018, there are six folders containing PDF files with NHANES response rate data and SAS Transport files for each of the investigation measurement factors. Following importing the primitive datasets into Python, data extraction and processing was essential to identify and classified variables. We generated a Github repository including the original NHANES files, and the final dataset applied for constructing and evaluating the model.

The prediction model was trained and evaluated using data from the National Health and Nutrition Examination Survey (NHANES), which was gathered between 2007 and 2018. The purpose of developing this model was to evaluate the disease risk of hypertension using relevant risk factors in a representative sample of American adults aged 20 and older (n = 11,341). According to some studies [56, 62] related to high blood pressure, they all exclude people under the age of 20. The main reason is that the occurrence of hypertension at the age of 20 is mainly related to genetic factors. According to the American Heart Association's

definition of hypertension, which uses blood pressure as the dichotomous dependent variable in this study, hypertension is defined as having a systolic blood pressure that is more than or equal to 140 mmHg [62]. Following the cleaning of the data, we used only the records that included values that were not null. Table 4.1 presents the distribution of samples based on the type of hypertensive people, as well as the people's gender and race.

| **Category** | **Gender** | **Ethnicity** | **Number** |
|---|---|---|---|
| Without hypertension | Female | Mexican American | 353 |
| | | Other Hispanic | 312 |
| | | Non-Hispanic White | 1,761 |
| | | Non-Hispanic Black | 519 |
| | | Other Race | 208 |
| Without hypertension | Male | Mexican American | 667 |
| | | Other Hispanic | 460 |
| | | Non-Hispanic White | 2,171 |
| | | Non-Hispanic Black | 738 |
| | | Other Race | 428 |
| hypertension | Female | Mexican American | 107 |
| | | Other Hispanic | 108 |
| | | Non-Hispanic White | 629 |
| | | Non-Hispanic Black | 376 |
| | | Other Race | 50 |
| hypertension | Male | Mexican American | 362 |
| | | Other Hispanic | 239 |
| | | Non-Hispanic White | 1,024 |
| | | Non-Hispanic Black | 653 |
| | | Other Race | 176 |

**Table 4.1:** Number of people by hypertension category, gender and ethnicity.

We conducted literature research that have used machine learning techniques to predict the occurrence of hypertension among different populations to identify several risk factors, including demographic variables in terms of age [174, 53, 175, 176], gender [53, 175, 62], race [56, 177, 62], education [177, 175], examination data like body measures [174, 53, 177, 175, 178] and waist [176, 178], chronic diseases in terms of diabetes and kidney conditions [53, 56, 62] and lifestyle factors such as smoking cigarette use [175, 56, 62], alcohol use [177, 179], exercise [177, 175], diet [177, 175] and sleeping [174, 180]. We used hist charts to simply analyze the correlation between these features with hypertension as shown in Figure 4.3.

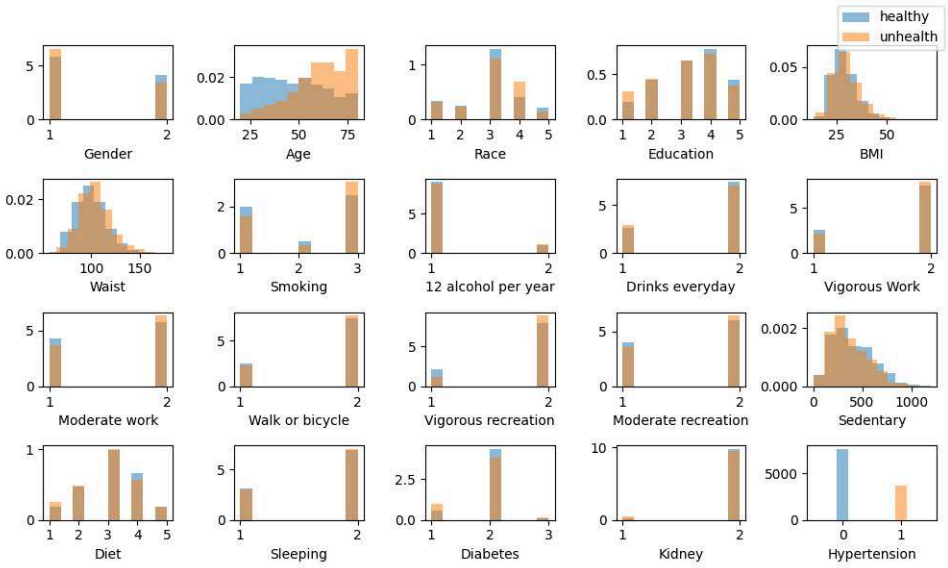



**Figure 4.3:** Distribution of hypertension for features

From Figure 4.3, individuals with hypertension in different cohorts have increased kidney disease, diabetic issues, and a notable relationship with unhealthy habits throughout follow-up. Although it has been shown that careful management of BMI can reduce the incidence of hypertension (López-Martínez et al., 2020), other factors such as age, race, education level, and lifestyle choices also influence the prevalence of hypertension. Meanwhile, the number of healthy and unhealthy people is imbalanced according to the last subplot. Therefore, based on the previous analysis, 19 features including age, gender, race, education level, BMI, waist, smoking, drinking, physical exercise, sleeping, diabetes, and kidney problems were chosen as input features. Table 4.2 and Table 4.3 show all the selected variables.

| **Variable Code** | **Variable Description** | **Code** | **Description** |
|---|---|---|---|
| RIAGENDR | Gender | 1 | Male |
| | | 2 | Female |
| RIDRETH1 | Race/Hispanic origin | 1 | Mexican American |
| | | 2 | Other Hispanic |
| | | 3 | Non-Hispanic White |
| | | 4 | Non-Hispanic Black |
| | | 5 | Other Race |
| DMDEDUC2 | Education level | 1 | Grade lower than ninth |
| | | 2 | 9-11th grade (Consists of 12th grade without a diploma) |
| | | 3 | High school graduate/GED or equivalent |
| | | 4 | University or AA degree |
| | | 5 | A university degree or higher |
| SMQ040 | Do you currently keep smoking? | 1 | Yes |

Continued on next page

| Variable Code | Variable Description | Code | Description |
|---|---|---|---|
| | | 2 | No |
| ALQ101 | A minimum of 12 alcoholic beverages each year? | 1 | Yes |
| | | 2 | No |
| ALQ151 | Have you ever had 4/5 or even more drinks each day? | 1 | Yes |
| | | 2 | No |
| PAQ605 | Vigorous work activity | 1 | Yes |
| | | 2 | No |
| PAQ620 | Moderately active work | 1 | Yes |
| | | 2 | No |
| PAQ635 | Walk or ride a bike | 1 | Yes |
| | | 2 | No |
| PAQ650 | Vigorous recreational activities | 1 | Yes |
| | | 2 | No |
| PAQ665 | Moderately active recreation | 1 | Yes |
| | | 2 | No |
| DBQ700 | How healthy is the diet | 1 | Excellent |
| | | 2 | Very good |
| | | 3 | Good |
| | | 4 | Fair |
| | | 5 | Poor |
| SLQ050 | Have you ever mentioned your trouble sleeping to a physician? | 1 | Yes |
| | | 2 | No |
| DIQ010 | Your physician informed you that you have diabetes. | 1 | Yes |
| | | 2 | No |
| | | 3 | Borderline |
| KIQ022 | Ever told you that your kidneys are weak and failing | 1 | Yes |

Continued on next page

| Variable Code | Variable Description | Code | Description |
|---|---|---|---|
| | | 2 | No |
| HYPCLASS | Systolic: Mean blood pressure (mmHg) | 1 | Non-Hypertensive |
| | | 2 | Hypertensive |

**Table 4.2:** Selected categorical variables in the NHANES dataset

| Variable Code | Variable Description | Mean | Standard |
|---|---|---|---|
| RIDAGEYR | Age at Screening Adjudicated | 1.36 | 17.03 |
| BMXBMI | Body Mass Index ($kg/m_2$) | 29.0 | 36.57 |
| BMXWAIST | Waist Circumference | 100.85 | 16.15 |
| PAD680 | Minutes sedentary activity | 358.18 | 19.80 |

**Table 4.3:** Selected numerical variables in the NHANES dataset

### 4.2.2 Performance Evaluation

The precision, specificity, recall (sensitivity), accuracy, F1-measure, and AUC are the metrics that are applied in this research to evaluate the models that are suggested and compared. To begin, these metrics will be described with the help of a confusion matrix, which can be seen in Figure 3.2. According to it, the number of instances of each type (true positive, true negative, false positive,

and false negative) is indicated by the letters TP, TN, FP, and FN, respectively. Table 4.4 explains how the confusion matrix is used to calculate six indicators.

| **Performance measure** | **Mathematical equation** | **Remark** |
|---|---|---|
| Precision | $\frac{TP}{TP+FP}$ | The fraction of true hypertension samples among the classified hypertension samples. |
| Specificity | $\frac{TN}{TN+FP}$ | The percentage of healthy samples that were accurately categorized. |
| Recall (Sensitivity) | $\frac{TP}{TP+FN}$ | Identifies the proportion of hypertension samples that have been correctly classified. |
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ | Calculates the overall proportion of samples that have been successfully categorized. |
| F1-measure | $\frac{2PrecisionRecall}{Precision+Recall}$ | The harmonic average of the value of recall and precision. |
| AUC | $\frac{1}{2} \times (\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$ | The diagnostic ability of a classifier system to distinguish between non-hypertension and hypertensive people. |

**Table 4.4:** The introduction of six performance indicators

### 4.2.3 Experimental Setup

This research mainly focuses on improving the AUC of the ensemble classification approach because it tells how much the model is capable of distinguishing between classes. For performance evaluation, firstly the proposed ensemble learning approach is compared with various individual learners such as multilayer perceptron (MLP) [55], k-nearest neighbors (KNN) [181], Decision Tree (DT) [182], support vector machine (SVM) [183], Gaussian Naive Bayes (Gaussian NB) [184] and Logistic Regression Model (LRM) [185], which are mostly utilized in the existing research on the diagnosis of hypertension. Secondly, the

proposed method is compared with six well-known ensemble learning methodologies namely bagging, boosting, and stacking. Specifically, random forest (RF) uses a bagging ensemble technic based on multiple decision trees, and Adaptive Boosting (AdaBoost), Extreme gradient boosting (XGBoost), and Light gradient boosting machine (LightGBM) are based on residual iterative tree. Further, two state-of-the-art staking ensemble models are used. The paper [186] developed a stacking-based evolutionary ensemble learning system 'NSGA-II-Stacking' for predicting the onset of Type-2 diabetes mellitus based on SVM and DT. Then, the paper [164] proposes an optimal stacking ensemble approach combining different learning algorithms, which selects meta-learners following a multi-objective evolutionary algorithm named non-dominated sorting genetic algorithms-II. We utilized the Standard Scaler approach to normalize the dataset first because KNN and SVM are easily affected by feature scale. All experiments were simulated on a machine with Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz 1.80 GHz, 8 GB RAM., Windows 10 64-bit O.S., and Python 3.8.6 environment.

### 4.2.4 Hyperparameter Optimization

The parameter adjustment range of all models is set to a commonly used range and the final setting of parameters is carried out by using Bayesian optimization [145]. Specifically, the study population (11,341) was split into a training dataset and a testing dataset. The training dataset was derived from a random sampling of 70% (7,939) of the extracted study population and the testing sampling of the remaining 30% (3,402) to evaluate the model on data sets with known

labels (ground truth) that were never used for training. Therefore, we employed Bayesian optimization and five cross-validations to search parameters using the training dataset, which is implemented by the hyperopt package [187] in Python. The maximum iterative time is set as 50. The hyperparameter space for models is shown in Table 4.5.

| **Model Name** | **Hyperparameter** | **Options/Range** | **Selected value** |
|---|---|---|---|
| MLP | hidden_layer_sizes | [(50,50,50),(50,100,50),(100,)] | (50,100,50) |
| | activation | ['tanh','relu'] | relu |
| | solver | ['sgd','adam'] | sgd |
| | alpha | ['constant','adaptive'] | constant |
| | learning_rate | [0.0001,0.01,0.05,0.1] | 0.1 |
| | max_iter | [*range(100,500,100)] | 300 |
| KNN | n_neighbors | ['uniform','distance'] | distance |
| | weights | [*range(1,15)] | 14 |
| DT | splitter | ['best','random'] | best |
| | criterion | ["gini","entropy"] | entropy |
| | max_depth | [*range(1,50,5)] | 5 |
| | min_samples_leaf | [*range(1,15)] | 11 |
| | class_weight | ['balanced',None] | balanced |
| SVM | kernel | ['linear','poly','rbf','sigmoid'] | rbf |
| | gamma | [0.001,0.01,0.1,1] | 0.01 |
| | C | [0.001,0.01,0.1,1,10,100,1000] | 1 |
| | class_weight | ['balanced',None] | None |
| LRM | solver | ['newton-cg','lbfgs','liblinear','sag','saga'] | liblinear |
| | penalty | ['l1','l2','elasticnet','none'] | l2 |
| | C | [0.001,0.01,0.1,1] | 0.1 |
| | class_weight | ['balanced',None] | None |
| | max_iter | [*range(100,800,100)] | 700 |
| RF | criterion | ["gini","entropy"] | entropy |

Continued on next page

| Model Name | Hyperparameter | Options/Range | Selected value |
|---|---|---|---|
| | max_depth | [*range(1,15),None] | None |
| | min_samples_leaf | [*range(1,50,5)] | 1 |
| | class_weight | ['balanced',None] | balanced |
| | n_estimators | [*range(100,500,100)] | 400 |
| AdaBoost | n_estimators | [*range(100,500,100)] | 400 |
| | learning_rate | [0.01,0.05,0.1,1] | 0.1 |
| XGBoost | max_depth | [*range(1,15),None] | 13 |
| | min_samples_leaf | [*range(1,50,5)] | 31 |
| | class_weight | ['balanced',None] | None |
| | n_estimators | [*range(100,500,100)] | 300 |
| | learning_rate | [0.01,0.05,0.1] | 0.05 |
| | subsample | uniform(0.3,1) | 0.7283 |
| LightGBM | max_depth | [*range(1,15),None] | 13 |
| | class_weight | ['balanced',None] | None |
| | n_estimators | [*range(100,500,100)] | 400 |
| | learning_rate | [0.01,0.05,0.1] | 0.1 |
| | subsample | uniform(0.3,1) | 0.3930 |
| | lambda_l1 | uniform(0,0.6) | 0.0435 |
| | lambda_l2 | [0,10,15,35,40] | 0 |

**Table 4.5:** Hyperparameter space for models

After optimization, the AUC values of models using default parameters and optimized parameters are shown in Figure 4.4.

In Figure 4.4, we found that while the performance of machine learning models like KNN, DT, and SVM is significantly impacted by varying parameter values, the identifying power of MLP, RF, and LRM utilizing various hyperparameters is comparable. In conclusion, hyperparameter optimization is necessary because it helps machine learning models find better parameters to improve performance. For example, the performance of the decision tree in this figure has

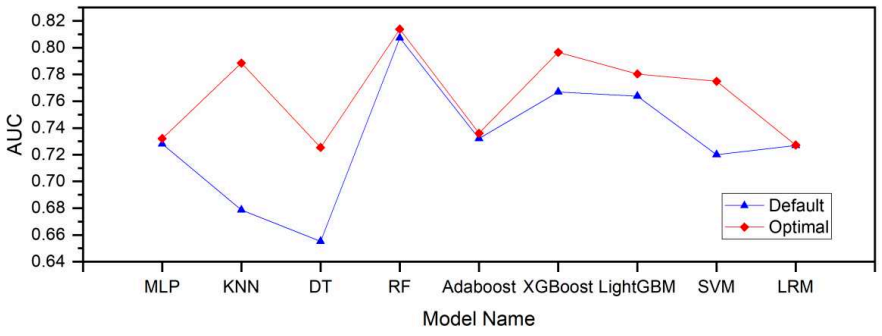



**Figure 4.4:** Comparison between default and optimized parameters

been significantly improved. It may be that the optimized parameters have improved its generalization ability, making the model perform better on untrained data.

### 4.2.5 Ensemble Model Construction

Another aspect that plays an important role in determining the accuracy of predictions is meta-learners. Growing the number of meta-learners could potentially enhance global generalization; however, an excessive number could result in overfitting. Meanwhile, the computing cost will increase proportionally with the amount of meta-learning done. Based on the proposed model selection approach, MoItMS, the procedure of model selection is shown in Figure 4.5.

According to the predicted values and real values, the objective values in terms of accuracy (E) and diversity (D) for each individual model can be calculated as shown in Figure 4.6.

We can see in Figure 4.6 that KNN has the best performance, whereas MLP and LRM have similar $E$ and $D$ values. Then, when we applied the proposed

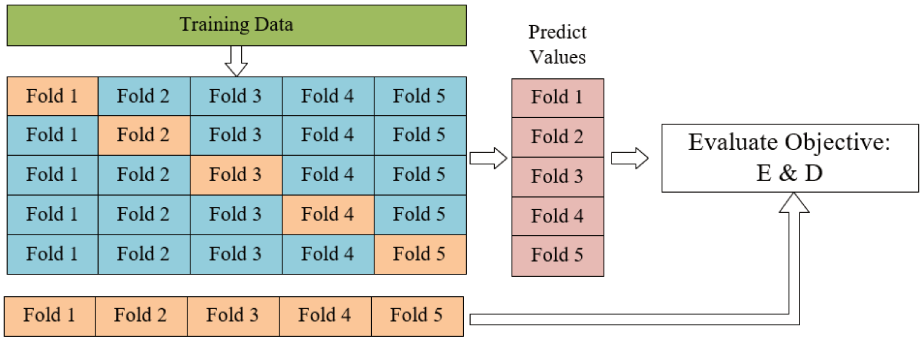



**Figure 4.5:** The procedure of model selection based on the accuracy (E) and diversity (D)

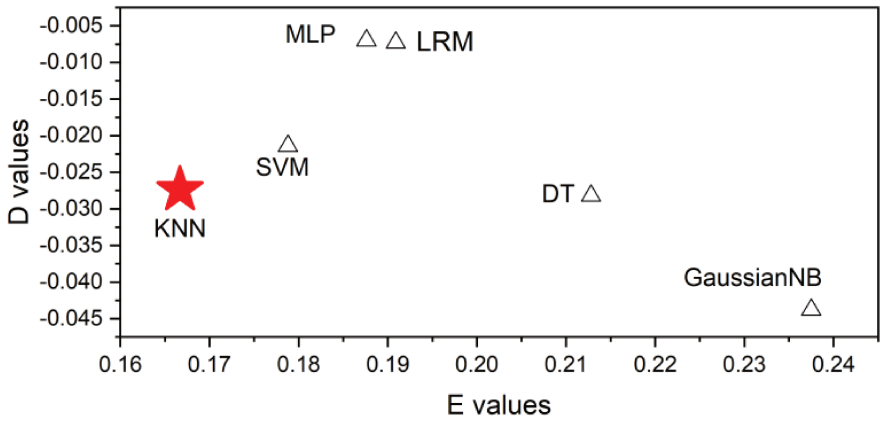



**Figure 4.6:** The objective values for models

MoItMS approach, the weight ($\lambda$) of diversity needs to be determined in equation (1). The search range of weight is denoted as $\lambda = 0.5, 1, 1.5, 2$, and the ensemble model's AUC values for the different weights are shown in Figure 4.7.

According to Figure 4.7. It is obvious that when $\lambda$ equals 1, the AUC value is greatest. Additionally, Figure 4.8 displays the AUC values for each iteration of the ensemble model when the weight is set to 1.

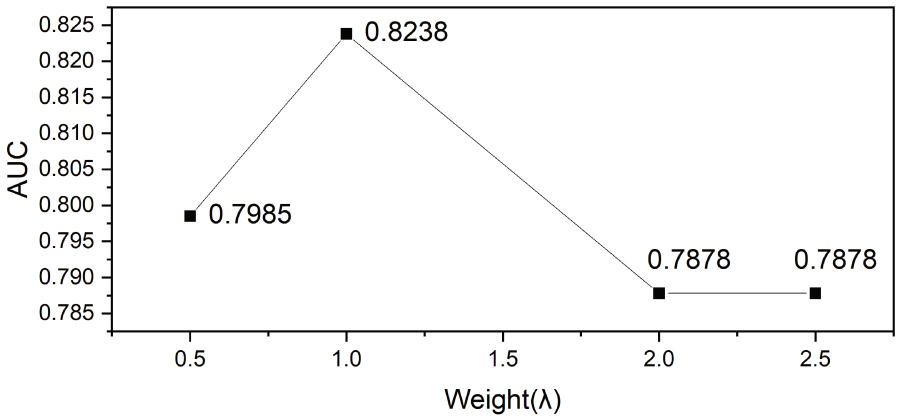



**Figure 4.7:** The ensemble model's AUC values for each weight

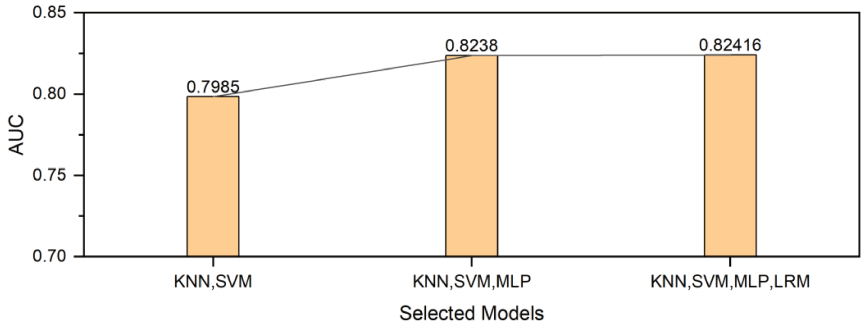



**Figure 4.8:** The ensemble model's AUC values for each iteration

As can be seen, the ensemble model with four models has the highest AUC, but its AUC value is similar to the ensemble model with three models. Concurrently, the amount of time spent computing rises in a steady and predictable manner due to the increased computational burden caused by the accuracy and diversity of computations performed inside the aggregative model. Therefore, the most suitable ensemble model is the aggregation of the three models. Specifically, KNN, SVM, and MLP are the three different meta-learners that are chosen by

the proposed method in accordance with the MoItMS methodology. Additionally, stacking is used in this paper for better fusion, and a neural network model with a hidden layer is used for the meta-classifier. This is due to the fact that the neural network model has the potential to produce, and that having one hidden layer can shorten the time that is consumed.

### 4.2.6 Model Evaluation

In this section, a comparative analysis of the suggested approach and thirteen other methods is carried out. The results of 20 separate simulations are summarized in Table 4.6, which compares the proposed stacking ensemble model against a total of six distinct individual models.

| **Individual Models Name** | **Precision** | **Recall** | **Accuracy** | **F1-measure** | **AUC** |
|---|---|---|---|---|---|
| MLP | 0.5637 | 0.4105 | 0.7002 | 0.4733 | 0.7383 |
| KNN | 0.6588 | 0.4994 | 0.7495 | 0.5679 | 0.8154 |
| DT | 0.4820 | **0.7522** | 0.6514 | 0.5872 | 0.7254 |
| SVM | 0.0 | 0.0 | 0.6702 | 0.0 | 0.7968 |
| Gaussian NB | 0.4989 | 0.3604 | 0.6697 | 0.4182 | 0.6940 |
| LRM | 0.5604 | 0.3600 | 0.6957 | 0.4382 | 0.7304 |
| Proposed Staking | **0.7113** | 0.5376 | **0.7682** | **0.6105** | **0.8420** |

**Table 4.6:** Performance comparison with individual classifiers

Considering the results of Table 4.6, it is evident that in the context of accuracy, the proposed methodology achieves a maximum AUC value of 0.8425 succeeded by individual learner KNN (0.8154). In addition, the proposed model showed significant improvement in the accuracy indicator (0.7682) compared with

the other six individual models. Although the recall value of the proposed model (0.5376) was low than DT (0.7522), it outperformed obviously than DT on the other four indicators. Further, the average performances of the proposed stacking ensemble model are compared with 6 ensemble models in Table 4.7. Here, two sophisticated stacking ensemble models are used as benchmarks in this study. [186] SVMs and DTs were used as the base learner, and the NSGA-II algorithm was used to combine models that were trained on different sub-datasets. In the paper [164], the NSGA-II algorithm was used to choose a model from a set of individual and tree-based ensemble models. In addition, voting is usually beneficial when aggregating a large number of base learners that attain comparable performance for similar work. As a result, an ensemble model based on Majority Voting is used as a benchmark against which the proposed stacking framework is measured.

| **Models Name** | **Base Learners** | **Ensemble Technic** | **Precision** | **Recall** | **Accuracy** | **F1-measure** | **AUC** |
|---|---|---|---|---|---|---|---|
| RF | DT | Bagging | 0.6991 | 0.4922 | 0.7626 | 0.5775 | 0.8306 |
| Adaboost | DT | Boosting | 0.5689 | 0.3658 | 0.6993 | 0.4451 | 0.7365 |
| XGBoost | DT | Boosting | 0.6564 | **0.5394** | 0.7549 | 0.5920 | 0.8102 |
| LightGBM | DT | Boosting | 0.6409 | 0.5181 | 0.7453 | 0.5729 | 0.7895 |
| [186] | SVM, DT | Stacking | 0.5508 | 0.4421 | 0.6967 | 0.4893 | 0.7335 |
| [164] | RF, XGBoost, LightGBM, MLP | Stacking | 0.6947 | 0.5108 | 0.7637 | 0.5871 | 0.8361 |
| Majority Voting | KNN, SVM, MLP | Majority Voting | 0.5582 | 0.3974 | 0.6971 | 0.4630 | 0.7359 |
| Proposed Staking | KNN, SVM, MLP | Stacking | **0.7113** | 0.5376 | **0.7682** | **0.6105** | **0.8420** |

**Table 4.7:** Performance comparison with other ensembles

Table 4.7 shows that our model (0.8420) had the best performance with AUCs, followed by the paper [164] (0.8361). Surprisingly, Majority Voting's AUCs (0.7359) is dismal, even worse than KNN's, implying that Majority Voting is not suitable as a simple ensemble approach in our study. On the other hand, the stacking architecture that we utilized possesses a substantial benefit in the sense that it is able to learn the values that are produced by each model. In terms of recall, our model achieved the highest value possible, which was 0.5376, followed by XGBoost (0.5394). The proposed technique achieves the highest value in terms of precision, which is 0.7113. This is followed by Random Forest, which achieves 0.6991, and the paper [164] achieves 0.6947. The specificities displayed by the paper [186] (0.5508) and Majority Voting was the most problematic (0.5582). The proposed strategy was able to obtain an average F1-measure that was 0.6105, making it the most successful method overall. In accuracy terms, the best performance was obtained from our model, followed by the paper [164] (0.7637) and Random Forest (0.7626). Additionally, our method achieves better performance than the paper [164] with a smaller number of models and has lower complexity in the process of model selection. The proposed approach's promising and competitive performance results demonstrated its superiority to the conventional stacking approach. In conclusion, in terms of prediction performance, the suggested stacking technique surpasses both the six individual and seven ensemble approaches.

Furthermore, a boxplot depicts the distribution of the data and is helpful in determining whether or not there are typical observations or outliers present

in the data. The boxplots of the five performance measures in terms of precision, recall, accuracy, F1-measure, and AUC obtained using a variety of models and the proposed ensemble method are depicted from Figure 4.9 to Figure 4.13, respectively. The outlier is shown by the sign "+" in each of the figures. Twenty iterations of each classifier are utilized in order to acquire the boxplot values for each metric. Additionally, two advanced methods [186, 164] are denoted as "Singh & Singh" and "Li et al.," in these figures.

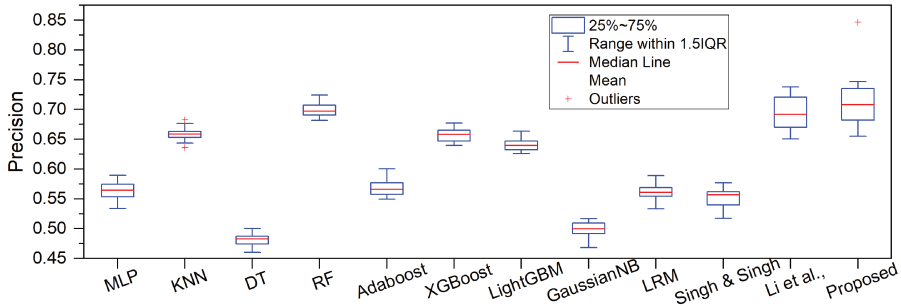



**Figure 4.9:** Boxplot of percentage precision

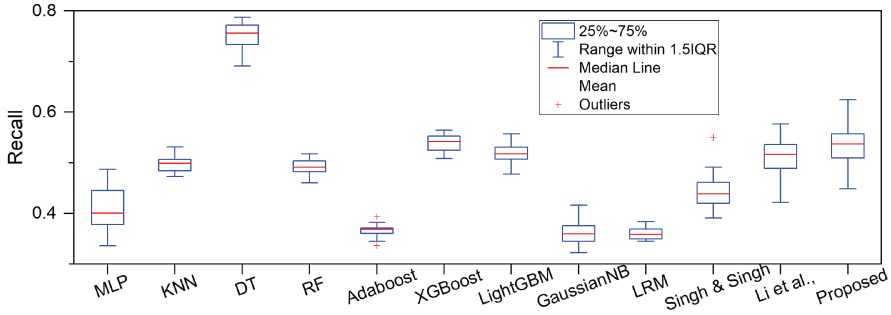



**Figure 4.10:** Boxplot of percentage recall

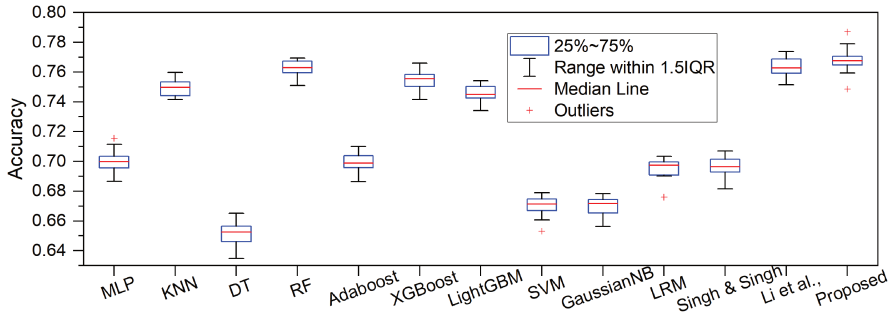



**Figure 4.11:** Boxplot of percentage accuracy

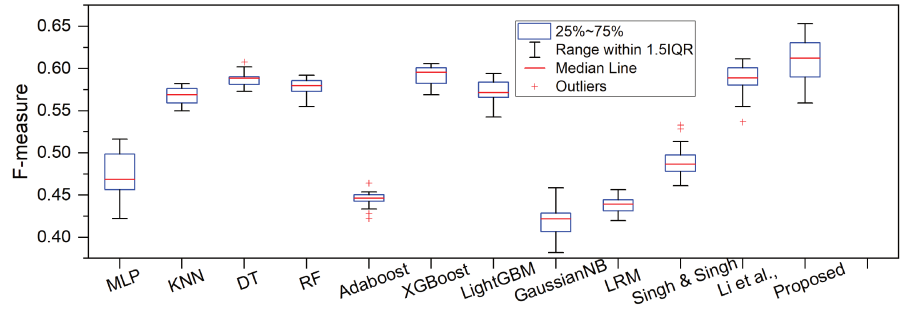



**Figure 4.12:** Boxplot of percentage F1-measure

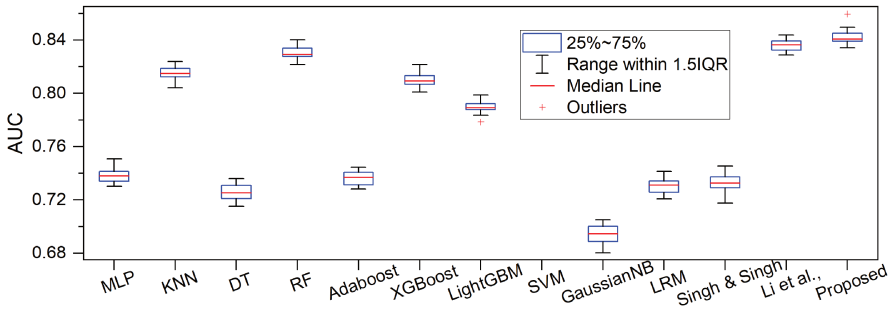



**Figure 4.13:** Boxplot of percentage AUCs

When comparing the precision distributions in Figure 4.9, it can be seen that the proposed methodology offers the highest precision value when compared

to the other methods. DT and Gaussian Naive Bayes Network (Gaussian NB) are the two approaches with the lowest precision values. When looking at the distributions of recall in Figure 4.10, it can be seen that DT achieved the highest recall value, followed by the proposed technique. Despite the fact that DT appeared to have the highest recall, it generated the least amount of precision and accuracy. The accuracy distributions are displayed in Figure 4.11 and indicate that the suggested stacking strategy achieves a much greater accuracy when compared to the other classifiers. These algorithms, including DT, SVM, and Gaussian NB, produce lower accuracy values. The ensemble learner [164] achieves the second-lowest accuracy of all the learners shown in this image. It is evident from the distributions of the F1-measure that are presented in Figure 4.12 that the strategy that has been proposed produces the highest F1-measure value. The AdaBoost, DT, and LRM techniques, on the other hand, produce solutions with lower F1-measure values. Finally, the area under the curve (AUC) comparisons of the proposed technique and the benchmark method are shown in Figure 4.13. As can be seen in this figure, the suggested method performed better than any of the other classifiers when it came to AUC. The paper [164] came in second, which suggests that the proposed stacking model performs better than the complex model. The greater AUC is largely attributable to the aggregation of the decision-making capabilities of the chosen base learners, which are then combined with the suitable meta-learner. Therefore, in terms of predicted precision, accuracy, F1-measure, and AUC, the suggested method fared better than all of the individual and ensemble approaches. The overall positive performance

of the suggested methodology may be valuable in assisting doctors in providing diagnoses that are more accurate and trustworthy, and it may have significant promise in the field of clinical hypertension diagnosis. In addition, the classification report generated by our model is included in Table 4.8 for the purpose of carrying out analysis in the clinical sense. Additionally, sensitivity and specificity can be determined using 4.4 and are displayed in Table 4.8 respectively.

| **True Positive (TP)** | **False Negative(FN)** | **False Positive(FP)** | **True Negative(TN)** | **Sensitivity** | **Specificity** |
|---|---|---|---|---|---|
| 623 | 536 | 253 | 1990 | 0.5376 | 0.8872 |

**Table 4.8:** Classification Report

Since its sensitivity is only 53.76%, the model proposed here may be ineffective as a healthcare diagnostic tool for detecting people who are genuinely hypertensive. However, the model's true negative rate (88.72%) suggests that it is successful in detecting those who are not hypertensive. We can also see that our model has a high negative predicted value of 1,990/2,526 (or 78.11%), demonstrating its suitability as a testing instrument. As well as it has provided a reference value for positive prediction in 623 out of 876 (or 71.11%), which demonstrates that it is superior to an inference drawn at random.

## 4.3 Extensive Approach Evaluation

The paper has shown that the classification capability of the model improved (AUC=0.8420) when applied to the input features19 features. In previous research, the results of artificial neural networks (AUC=0.77) were utilized when

applied to the input features of gender, race, BMI, age, smoking, kidney conditions, and diabetes. In order to further explore the performance of our proposed approach, we conduced an experiment on the same dataset [62] with the previous research. According to the proposed approach for model selection, MLP, LRM, and Gaussian NB models are employed as base models in level-0. Six machine learning algorithms in the paper [62] were identified and compared, including decision jungle, logistic regression, support vector machine, boosted decision tree, Bayes point machine, and artificial neural network. Among them, parameters of MLP, LRM, and Gaussian NB from the paper [56], and parameters of other models are optimized by the Bayesian Optimization algorithm. The experiment results are shown in Table 4.9.

| **Models Name** | **Precision** | **Recall** | **Accuracy** | **F1-measure** | **AUC** |
|---|---|---|---|---|---|
| SVM | 0.59 | 0.464 | 0.737 | 0.464 | 0.759 |
| DJ | 0.581 | 0.453 | 0.734 | 0.453 | 0.769 |
| BDT | 0.564 | 0.462 | 0.729 | 0.462 | 0.765 |
| BPM | 0.583 | 0.456 | 0.735 | 0.456 | 0.763 |
| LR | 0.589 | 0.465 | 0.737 | 0.465 | 0.764 |
| ANN | 0.578 | 0.474 | 0.732 | 0.474 | 0.770 |
| Proposed Stacking | **0.592** | **0.490** | **0.745** | **0.536** | **0.788** |

**Table 4.9:** Classification methods comparison

The findings of a comparison of six distinct approaches with our suggested method are presented in Table 4.9. In terms of predictive precision, recall, accuracy, f1-measure, and area under the curve (AUC), we discovered that the proposed approach outperformed all other methods. Moreover, based on the f1-measure, our model scored the highest attainable value, which was 0.536, followed

by the Artificial Neural Network achieved 0.474. This is a significant improvement. Furthermore, our research considered lifestyle factors compared with the previous research[56, 62]. So as to explore the effect of lifestyle factors on hypertension prediction, a sub-dataset without lifestyle factors is used. The input features are age, gender, race, education, BMI, waist, diabetes, and kidney. The experiment results are shown in Table 4.10.

| **Datasets** | **Precision** | **Recall** | **Accuracy** | **F1-measure** | **AUC** |
|---|---|---|---|---|---|
| Dataset with lifestyle features | **0.7113** | **0.5376** | **0.7682** | **0.6105** | **0.8420** |
| Dataset without lifestyle features | 0.7104 | 0.4956 | 0.7668 | 0.5834 | 0.8409 |

**Table 4.10:** Comparing the impact of lifestyle factors in hypertension prediction

The experimental results show that after removing lifestyle characteristics, the prediction performance, including precision, accuracy, and AUC values, only slightly dropped, while recall and F1-measure decreased by 4.2% and 2.71%, respectively. As demonstrated in 4.3, the four characteristics of gender, age, education level, and obesity have strong discriminatory power for hypertension in our study. Furthermore, the model's capacity to correctly hypertensive samples is degrading, as evidenced by the fall in recall terms. In practice, a model with a better hypertension discrimination performance is preferable. Despite the slight performance gain, we still suggest integrating lifestyle features in the model because they can improve the model's performance while also assisting in the analysis of the causes of the patient's condition.

## 4.4 Summary

Various categorization algorithms for the early detection of lifestyle-related diseases have been presented in recent years. One of the current study areas is selecting an acceptable methodology that strikes a compromise between efficiency and implementation complexity. According to the reports of the National Health and Nutrition Examination Survey (NHANES), the prevalence of hypertension in the adult population of the United States is high and has been rising over the past few years. We initially devised a Multi-objective Iterative Model Selection (MoItMS) strategy to simultaneously maximize the ensemble model diversity and the accuracy of meta-learners in this work. Subsequently, a stacking-based aggregative method for accurately classifying the data of hypertension patients was created. The proposed model uses three distinct types of learners namely, KNN, SVM, and MLP, as its basic learners. Each of these models is trained using cross-validation to ensure accuracy. The level-1 data is comprised of the predictions made on training samples in addition to the actual labels, both of which are utilized in the process of training the meta-learner. After that, the meta-learner is used to make predictions regarding the testing samples. The effectiveness of the proposed ensemble technique is evaluated with reference to both individual and ensemble models, which serve as baseline models. The comparative findings reveal that the proposed model performs better than the baseline individual and ensemble models according to five specified evaluation measures, such as accuracy, precision, recall, F1-measure, and AUC value. These metrics include

accuracy, precision, and recall. In addition, we assessed the suggested stacking structure by employing hypertension datasets that included gender, race, BMI, age, smoking, kidney problems, and diabetes. According to the findings of the experiment, the proposed method performs better than the previous studies on all five of the evaluation measures that were used. Finally, we evaluated the effect of lifestyle factors on the classification performance for hypertension, and we found that lifestyle factors can help the model discriminate hypertensive samples from normal samples. In future studies, a more in-depth examination and screening of features will be considered. On the other hand, in order to verify the proposed framework, hypertension can be predicted using a variety of data sets, including those with various features and risk factors.

# Chapter 5. A Case Study for A Lifestyle-Related Disease

## 5.1 Data Source

This study used real medical data gathered during a hospital health check-up in Nanjing, China. This dataset is from 2012 to 2022. All subjects in the study gave informed consent to the use of the data, and all sensitive information about the subjects was removed from the original dataset. In this real case study, hypertension is an example of a lifestyle-related disease because it is really common in our daily life. First, we removed 23 records who were 20 years of age or younger. The remaining data comprised 32,784 instances and 65 attributes. Specifically, there are 41 features including age, gender, heart rate (HR), height, weight, waist circumference (WC), body massive index (BMI), hemoglobin (HB), white blood cell (WBC), platelets (PL), urinary protein (UP), Urinary sugar (US), Urinary ketones (UK), Urinary occult blood (UOB), blood sugar (BS), alanine aminotransferase (ALT), aspartate transaminase (AT), Total Bilirubin (TB), Creatinine (CR), BU (Blood Urea), Total cholesterol (TC), triglycerides (TG), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), right systolic blood pressure (right_SBP), right diastolic blood pressure (right_DBP), left systolic blood pressure (left_SBP), left diastolic blood pressure (left_DBP), exercise frequency (L_EF), exercise year(L_EY), exercise time

(L_ET), smoking (L_S), smoking quantity(L_SQ), smoking age(L_SA), drinking frequency(L_SQ), drinking quantity(L_DQ), drinking age (L_DA), diet balance (D_BD), diet hobby (D_DH), Atherosclerosis (As), fat liver (FL), hypertension (HTN). In addition, there are 24 symptoms including blurred vision (S_BV), dizziness (S_Di), polydipsia (S_ polydipsia), polyuria (S_ polyuria), vertigo (S_vertigo), headache (S_HA), joint swelling and pain (S_joint_SP), numb hands and feet (S_numb_HF), tinnitus (S_tinnitus), constipation (S_constipation), chest tightness (S_CT), palpitations (S_palpitations), nausea and vomiting (S_NV), chest pain (S_CP), chronic cough (S_CC), fatigue (S_fatigue), sputum production (S_SP), diarrhea (S_diarrhea), weight loss (S_WL), urgency (S_urgency), dyspnea (S_dyspnea), painful urination (S_PU), breast pain (S_BP). Meanwhile, there are 18,936 males (57.75%) and 13,848 females (42.24%) in the dataset, with an age of 63.88±9.27.

## 5.2 Missing Value Analysis and Processing

First, the missing value module can automatically calculate the missing rate of each dimension in the dataset. First, the missing value module automatically calculates the missing rate of each dimension in the data set. Specifically, the overall missing rate in our case is 13.36%. Subsequently, the missing value module automatically analyzes the absence of the missing rate of the features in the data set, as shown in Figure 5.1.

Figure 5.1 shows that some features' missing rate exceeds the 0.8 cutoff point, which means that 80% of their values are lost. Because of this, we exclude these features, which include L_SQ, L_SA, L_DQ, and L_DA. Following that, as

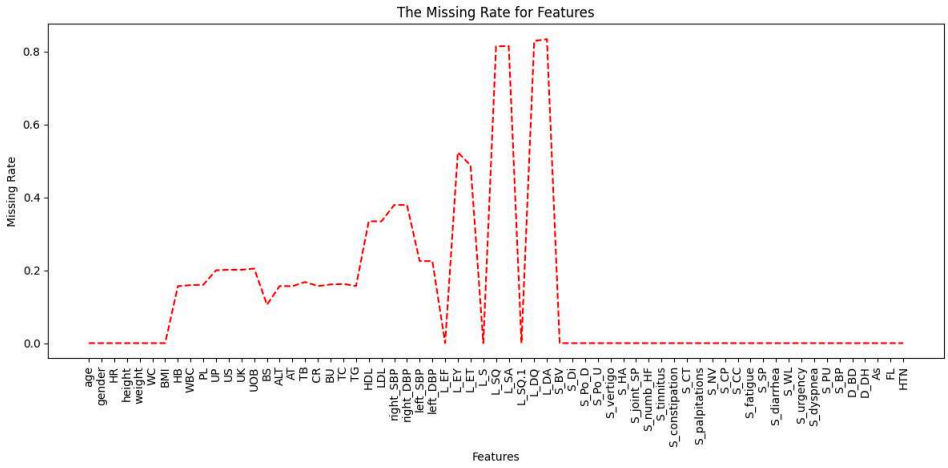



**Figure 5.1:** Missing rate of features in the case study

seen in Figure 5.2, the missing value module examines the absence of instances in the dataset. It is important to note that it is impossible to display the missing rate for each instance of a big data set, such as the more than 30,000 records in our case. In order to illustrate the distribution of each missing rate segment of the instance, the missing value module employs segmented statistics.

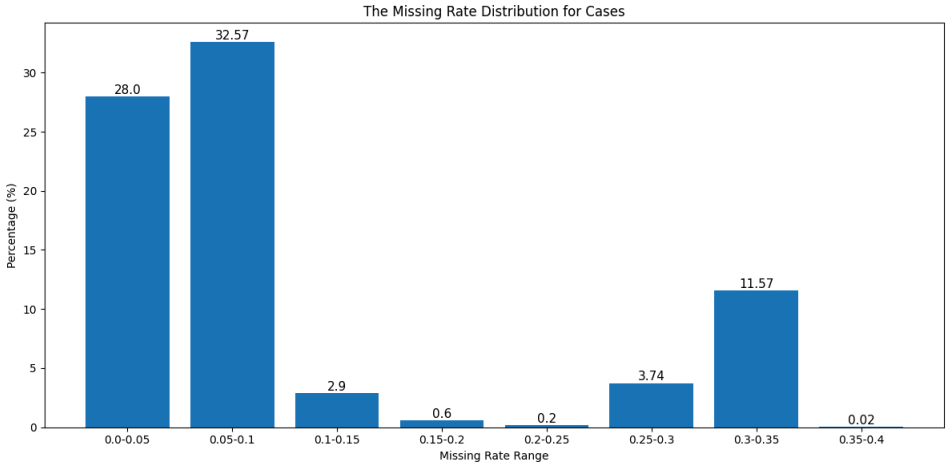



**Figure 5.2:** Segmented statistics of the missing rate of instances in the case study

Figure 5.2 shows that 28% of the dataset's instances have less than 10% of their values missing, while 32.57 of them have missing values between 10% and 20%. Less than 0.02% of the instances lost more than 35% of the values at the same moment. Overall, no instance's portion of the dataset is missing by more than 50%, hence no instance is disregarded. The missing value module also employs the distribution map and hot map of missing values for auxiliary analysis to examine the missing mechanism and missing mode of missing values, as shown in Figure 5.3 and Figure 5.4.

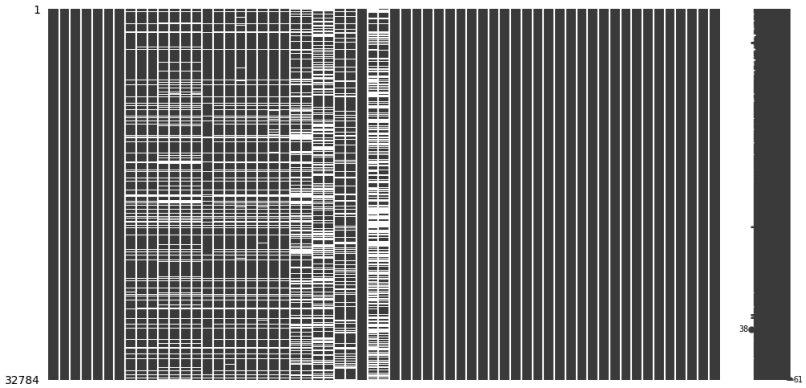



**Figure 5.3:** Distribution of missing values in the case study

Missing values are mainly distributed discretely in various measured features, as seen in Figure 5.3. At the same time, it can be shown that several features, such as ALT and HB, have a significant relationship according to the missing value heat map (Figure 5.4). It is not advised to delete the missing value model of the missing values in our data set directly since it is not missing completely at random (MCAR). The missing pattern in our case data is non-

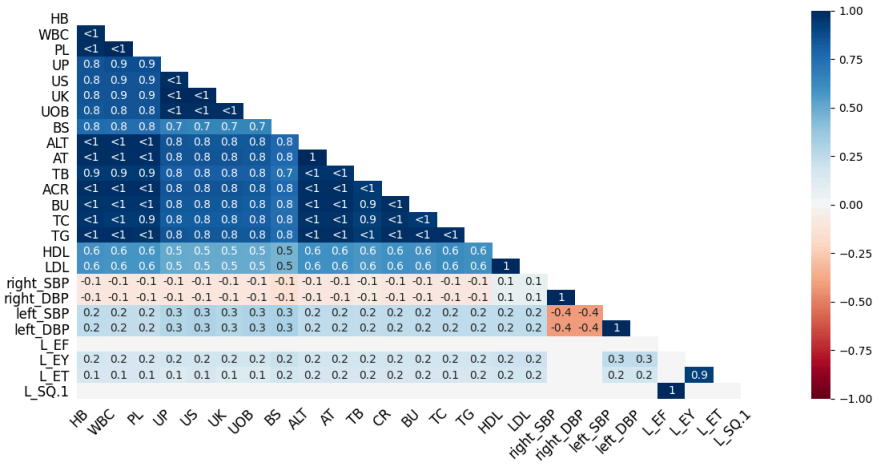



**Figure 5.4:** Hot map of missing values in the case study

monotonic, which is also supported by the distribution plot of missing values. The findings of the missing value analysis show that, even after eliminating some features with 80% missing values, the data set still contains 8.84% missing values. We examine the imbalance rate of categorical features with missing values in order to effectively handle these missing values. By dividing the number of classes with the most values in the feature by the number of classes with the fewest values, the imbalance ratio is determined. As seen in Figure 5.5, the imbalance rate analysis is carried out on the case's categorical features with missing values.

UP, UA, UK, and UOB are a few examples of categorical features with missing values that are noticeably uneven after looking at Figure 5.5. The UOB has the lowest imbalance rate of all of them at 7.67%. The proposed SncALWRFI imputation method was used to impute missing data based on the previous analysis. Pair deletion (PD), MEAN, KNNI, and MissForest missing value processing techniques were employed in comparison to examining the effects of the proposed

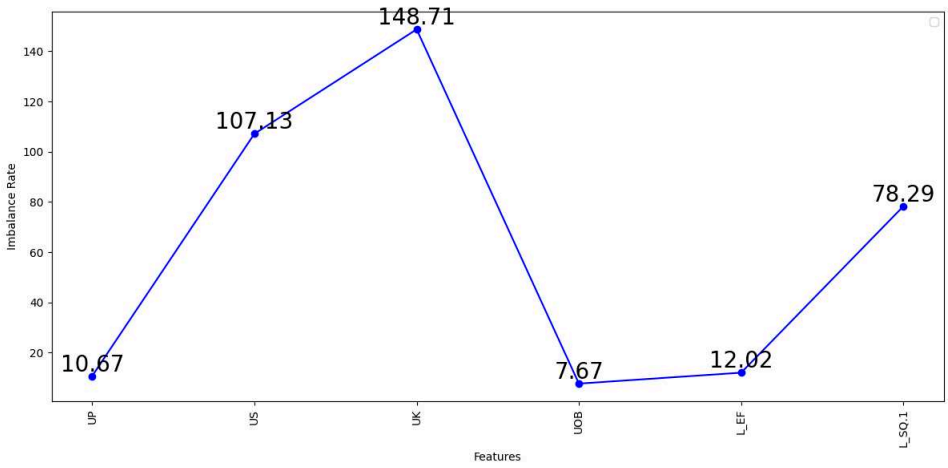



**Figure 5.5:** Imbalance rate analysis of categorical features in the case study

imputation approach on the performance of lifestyle-related disease prediction. Because 80% of the instances contain missing values, the complete case analysis (CCA) approach is not employed because it is impossible to delete instances with missing values.

Additionally, to ensure fairness, default parameters are chosen for datasets processed by various missing value methods, along with RF, LGBM, and LRM being used as predictive models for diseases connected to lifestyle. In more detail, the data is split into two sets: a training data set, which comprises 70% of the data, and a testing data set, which contains 30% of the data. The training data set is used to create a missing value imputation model, and the test data set is used to assess the model's effectiveness. We compare performance using AUC as a performance indicator. The experiment was carried out 20 times, and Table 5.1 displays the average outcomes.

| Methods | PD | MEAN | KNNI | MissForest | SncALWRFI |
|---|---|---|---|---|---|
| RF | 75.02 | 80.07 | 81.10 | 82.72 | **83.88** |
| LGBM | 75.98 | 81.46 | 82.92 | 83.94 | **84.83** |
| LRM | 71.08 | 72.19 | 71.91 | 72.20 | **73.31** |

**Table 5.1:** Prediction results of different processing methods for missing values in the case study

The maximum prediction result of 75.98 is obtained in the LGBM model, according to experimental results, while removing features with missing values yields the lowest prediction results. However, our proposed approach performs at its best, achieving an average ideal value of 84.83 in the LGBM model. Therefore, as the output data for the missing value module, we will ultimately select the data set without missing values that was processed using the suggested SncALWRFI approach.

## 5.3 Feature Selection Based on Feature Importance

The highly accurate and robust random forest-based feature selection (RF_FS) method was introduced in 5.1.2. In the feature selection module, specifically, the data without missing values preprocessed by the missing value module will be input, followed by the use of RF_FS to analyze the importance of features, and finally the selection of the data set containing only key features in accordance with the ranking of feature importance. A predictive model for LRDs was cre-

ated using an experimental dataset. Initially, there were 65 features in our case, but since 4 of them (L_SQ, L_SA, L_DQ, and L_DA) were 80% absent from the dataset, they were excluded and the remaining 61 features were input into the feature selection module. When calculating feature importance, the result will be rounded to 3 decimal places. The final output of RF_FS is shown in Figure 5.6.

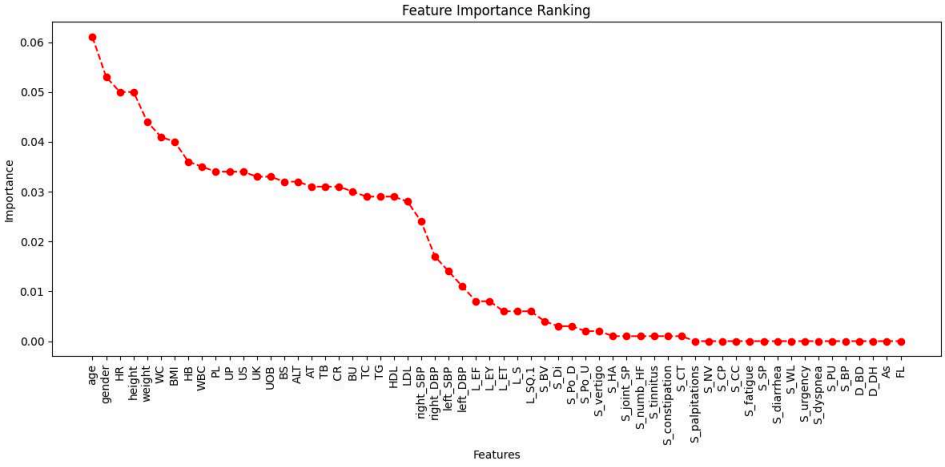



**Figure 5.6:** Ranking of feature importance in the case study

The top-N important features or all features with importance greater than 0 can be chosen once the calculation of feature importance is complete. In order to keep as many features as possible, the feature selection module selects according to the important threshold of the feature, that is, the features with importance of more than 0 are picked, and 16 features are then discarded. The final experimental dataset will have 32,784 instances and 45 features. We use the same three prediction models and conduct 20 runs to confirm the impact of feature selection strategies on LRDs' prediction outcomes. The Table 5.2 below

displays the average AUC results obtained from 20 runs using various prediction models.

| **Methods** | **RF** | **LGBM** | **LRM** |
|---|---|---|---|
| Non - Feature Selection | 83.88 | 84.83 | 73.31 |
| Random Forest Feature Selection | **84.17** | **85.28** | **73.89** |

**Table 5.2:** Prediction results of feature selection in the case study

The experimental results demonstrate that feature selection increased the performance of the three prediction models, demonstrating that the feature selection method based on random forest can increase the accuracy of LRDs prediction after removing some features with low importance.

## 5.4 The Construction of LRDs Ensemble Prediction Model

After analysis based on key features, the dataset with key features will be utilized to create a strong ensemble LRD predictive model. The final LRDs prediction model will be combined from candidate models including multilayer perceptron (MLP), K-Nearest Neighbors (KNN), Decision Tree (DT), support vector machine (SVM), Gaussian Naive Bayes Network (Gaussian NB), Logistic Regression Model (LRM) model, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Random Forest (RF). Three steps make up the model construction: ensemble model construction, hyperparameter optimization, and model evaluation. The disease prediction module will first

automatically adjust the hyperparameters of each individual model in order to improve performance. In Chapter 4.2.4, Table 4.5 provides a description of the parameter space. Each model will receive the ideal set of parameters following the Bayesian optimization procedure. The disease prediction module then uses the proposed MoItMS technique to select a suitable model for the ensemble case data. Figure 5.7 depicts the model's integration procedure.

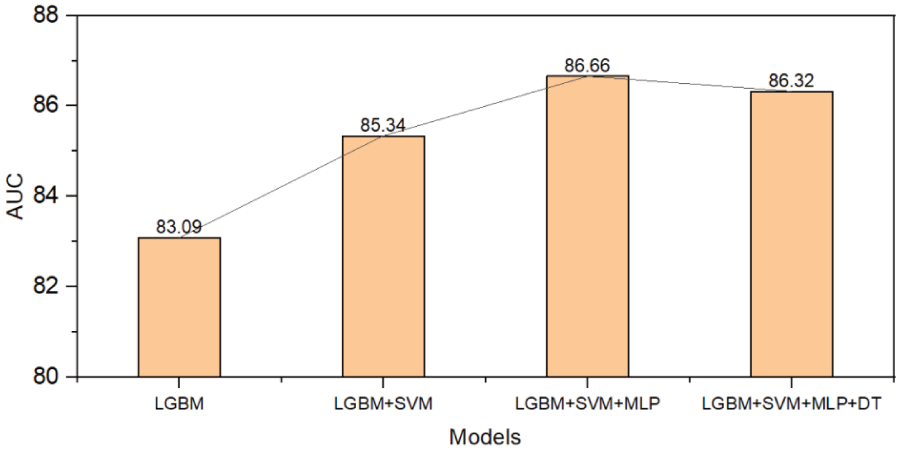



**Figure 5.7:** The generation process of the ensemble model

Figure 5.7 above shows the construction process of the integrated model. The LGBM model is selected at the beginning, and then more models are iteratively selected to be added to the integrated model according to the accuracy and diversity of the model. Due to the increased computational burden resulting from the accuracy and diversity calculations performed in the aggregated model, the amount of time spent on the calculation increases in a steady and predictable manner. Therefore, the most suitable ensemble model is the aggregation of the three models. Finally, the disease prediction module will use the six-dimensional

model capability map to automatically and visually evaluate the performance difference between the generated integrated model and the various sub-models that make up the model, as shown in Figure 5.8.

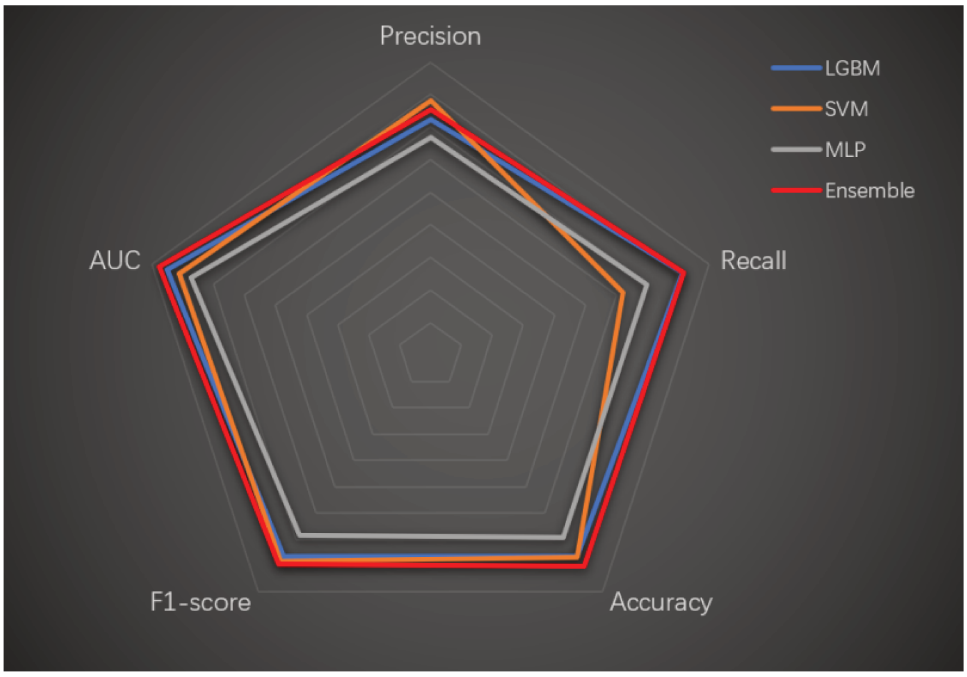



**Figure 5.8:** Ensemble model evaluation

According to Figure 5.8, it can be seen that the constructed integrated model presents the best performance in the capability chart. And the value of the most important AUC index is 87.54, which shows that the model has a high discrimination ability and has practical application significance. Finally, in order to further demonstrate the changes in model prediction performance after each module is processed, we compare and display them in Table 5.3.

| **Methods** | **RF** | **LGBM** | **LRM** | **Ensemble Model** |
|---|---|---|---|---|
| Original Data (MEAN) | 80.07 | 81.46 | 72.19 | - |
| Missing value module | 83.88 | 84.83 | 73.31 | - |
| Feature selection module | 84.17 | 85.28 | 73.89 | - |
| Disease prediction model | - | - | - | **87.54** |

**Table 5.3:** Prediction results of each module in the case study

## 5.5 Data Flow of the Prediction Framework

The data flow through the forecasting framework is then examined. In particular, the raw data will be appropriately processed in the proposed prediction framework and utilized to identify essential features and build core models, such as a missing value imputation model and an ensemble prediction model for LRDs. The prediction framework has three primary data processing components, which we previously analyzed:

1) The original data is converted into data without missing values and available and robust imputation models for missing values in the missing value module after some features and instances are removed and the null values are filled with the proper missing value processing method.

2) The feature selection module selects crucial features for lifestyle-related diseases using advanced feature selection techniques based on machine learning

and then turns the data into core experimental data for creating models for the prediction of lifestyle-related diseases.

3) The disease prediction module separates the training and test data sets, builds an integrated prediction model using the training data, assesses the usability of the prediction model using the test data, and finally transforms the data into a useable prediction model.

The data flow diagram of the proposed prediction framework is represented as Figure 5.9.

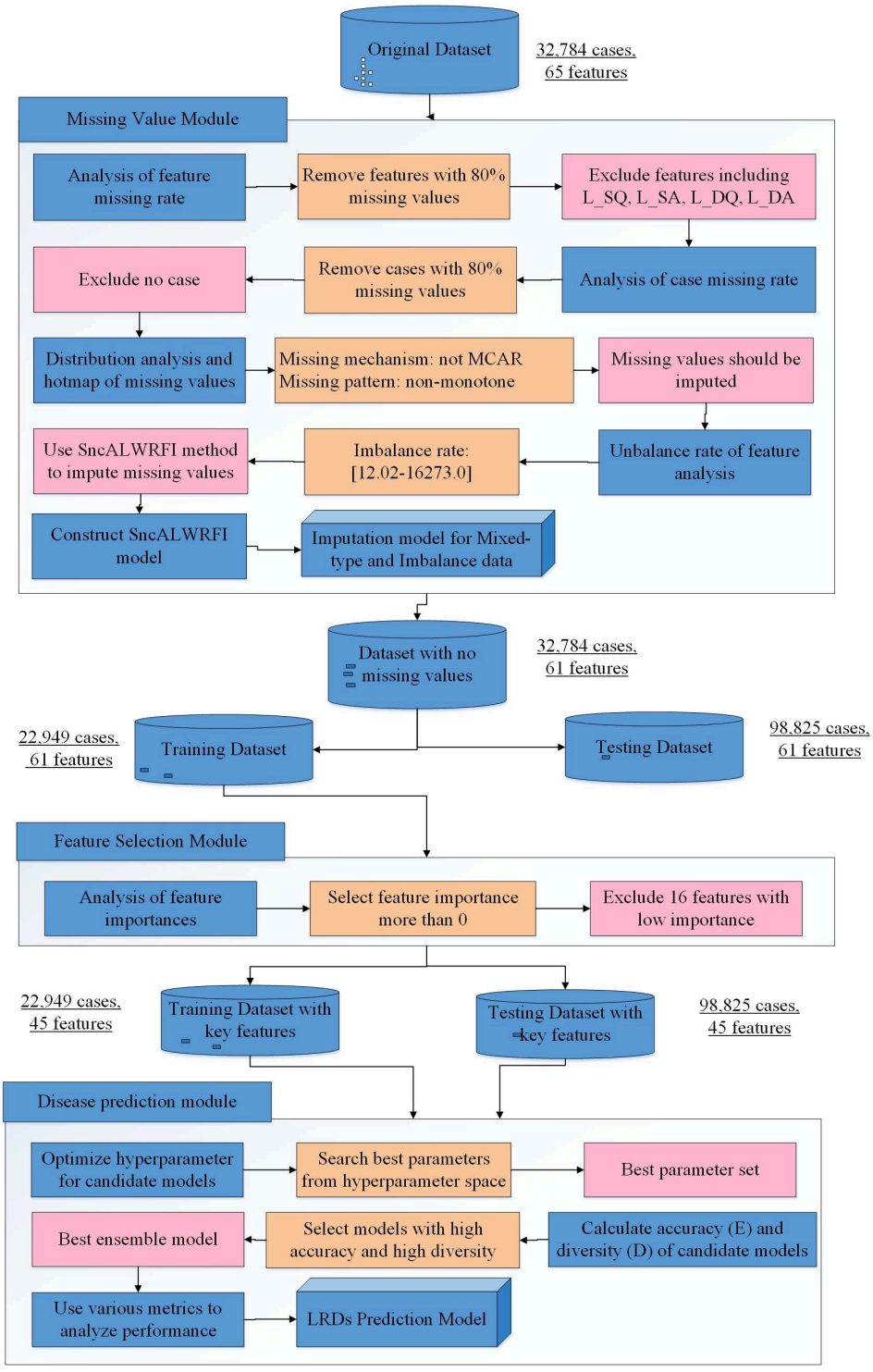



**Figure 5.9:** The data flow diagram of the proposed prediction framework

## 5.6 A Simple Application Scenario

We then present a simple scenario application for the proposed framework. Similar to the previous example, we will utilize the constructed prediction framework to forecast the probability of hypertensive diseases for a group of 10 new people who have had health examinations. Figure 5.10. displays the data for these people.

```
---------------Original Data-------------
   age  gender  HR  height  weight    WC  ...  S_PU  S_BP  D_BD  D_DH   As   FL
0   60     1.0  80     NaN    76.5  94.0  ...   0.0   0.0   0.0   0.0  0.0  1.0
1   69     0.0  71   158.0    57.0  78.0  ...   0.0   0.0   0.0   0.0  0.0  0.0
2   71     1.0  78   170.0    65.0  84.0  ...   0.0   0.0   0.0   0.0  0.0  0.0
3   78     1.0  58   166.5    65.3  89.0  ...   0.0   0.0   0.0   0.0  0.0  0.0
4   77     0.0  65   148.0    64.0  91.0  ...   0.0   0.0   1.0   0.0  0.0  1.0
5   79     0.0  59     NaN    67.3  90.5  ...   0.0   0.0   0.0   0.0  0.0  0.0
6   64     0.0  80   163.0    74.9  99.2  ...   0.0   0.0   0.0   0.0  0.0  0.0
7   48     NaN  77   165.0    60.0  70.0  ...   NaN   NaN   NaN   NaN  NaN  NaN
8   73     1.0  78     NaN    68.0  94.1  ...   NaN   NaN   NaN   NaN  NaN  NaN
9   68     0.0  64   147.0    38.0  63.0  ...   0.0   0.0   1.0   0.0  0.0  0.0

[10 rows x 64 columns]
```

**Figure 5.10:** The data for 10 people in the simple application scenario

First, the data is input into the feature selection module, and the execution result is shown in Figure 5.11.

```
------------Feature Selection Module-----------------
Exclude 16 features:
S_palpitations, S_NV, S_CP, S_CC, S_fatigue, S_SP, S_diarrhea, S_WL,
S_urgency, S_dyspnea, S_PU, S_BP, D_BD, D_DH, As, FL

-------------Data after feature selection module-------------
   age  gender  HR  height  ...  S_numb_HF  S_tinnitus  S_constipation  S_CT
0   60     1.0  80   163.1  ...        0.0         0.0             0.0   0.0
1   69     0.0  71   158.0  ...        0.0         0.0             0.0   0.0
2   71     1.0  78   170.0  ...        0.0         0.0             0.0   0.0
3   78     1.0  58   166.5  ...        0.0         0.0             0.0   0.0
4   77     0.0  65   148.0  ...        0.0         0.0             0.0   0.0
5   79     0.0  59   148.0  ...        1.0         0.0             1.0   0.0
6   64     0.0  80   163.0  ...        1.0         0.0             0.0   0.0
7   48     NaN  77   165.0  ...        NaN         NaN             NaN   NaN
8   73     1.0  78     NaN  ...        NaN         NaN             NaN   NaN
9   68     0.0  64   147.0  ...        0.0         0.0             0.0   0.0
```

**Figure 5.11:** The execution result of feature selection module in the simple application scenario

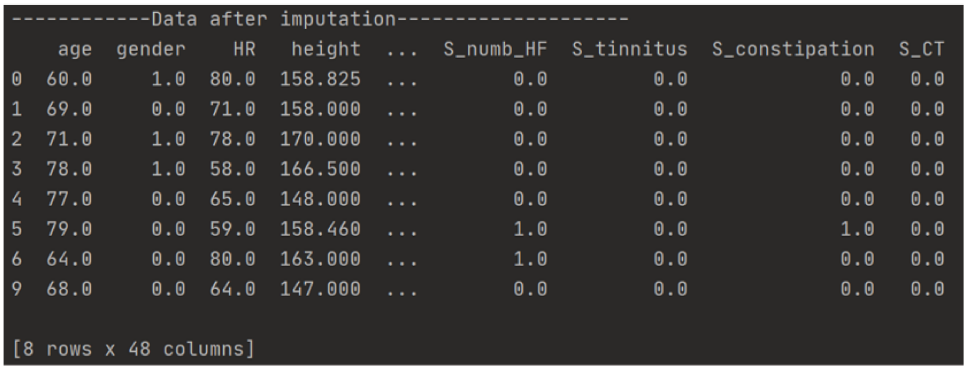
Next, input the data after feature selection into the missing value module, and the result is shown in Figure 5.12.

```
------------Missing Values Module-----------------
   age  gender  HR  height  ...  S_tinnitus  S_constipation  S_CT  Missing Rate (%)
0   60     1.0  80     NaN  ...         0.0             0.0   0.0             4.17
1   69     0.0  71   158.0  ...         0.0             0.0   0.0            16.67
2   71     1.0  78   170.0  ...         0.0             0.0   0.0            22.92
3   78     1.0  58   166.5  ...         0.0             0.0   0.0             4.17
4   77     0.0  65   148.0  ...         0.0             0.0   0.0            16.67
5   79     0.0  59     NaN  ...         0.0             1.0   0.0            14.58
6   64     0.0  80   163.0  ...         0.0             0.0   0.0             8.33
7   48     NaN  77   165.0  ...         NaN             NaN   NaN            87.50
8   73     1.0  78     NaN  ...         NaN             NaN   NaN            87.50
9   68     0.0  64   147.0  ...         0.0             0.0   0.0            22.92

[10 rows x 49 columns]
Two cases (ID=7, 8) will be excluded because their missing rate more than 85%.
Eight cases contain missing values and will be imputed.
```

**Figure 5.12:** The result of missing value module in the simple application scenario

According to the analysis results of the missing value module, two records (ID=7,8) will be excluded because they are missing more than 85% of the values. In addition, there are 8 records including missing values, which need to be imputed using the model. The imputed data are shown in Figure 5.13.

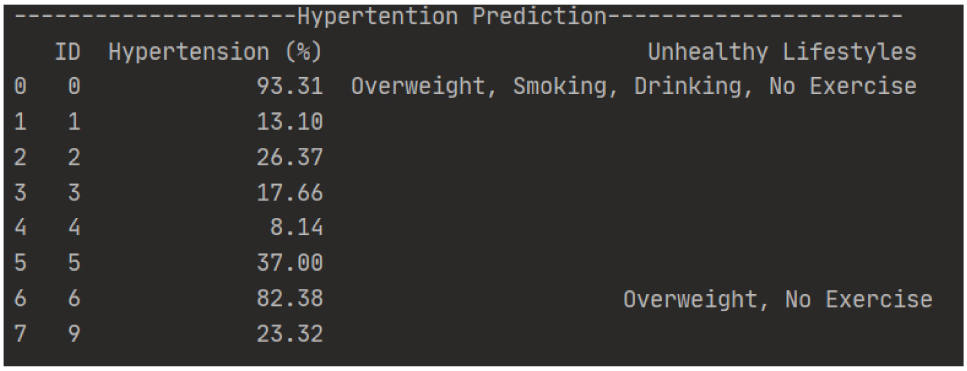


```
-----------Data after imputation--------------------
    age  gender    HR   height  ...  S_numb_HF  S_tinnitus  S_constipation  S_CT
0  60.0     1.0  80.0  158.825  ...        0.0         0.0             0.0   0.0
1  69.0     0.0  71.0  158.000  ...        0.0         0.0             0.0   0.0
2  71.0     1.0  78.0  170.000  ...        0.0         0.0             0.0   0.0
3  78.0     1.0  58.0  166.500  ...        0.0         0.0             0.0   0.0
4  77.0     0.0  65.0  148.000  ...        0.0         0.0             0.0   0.0
5  79.0     0.0  59.0  158.460  ...        1.0         0.0             1.0   0.0
6  64.0     0.0  80.0  163.000  ...        1.0         0.0             0.0   0.0
9  68.0     0.0  64.0  147.000  ...        0.0         0.0             0.0   0.0

[8 rows x 48 columns]
```

**Figure 5.13:** The imputed data in the simple application scenario

Finally, we input the processed data into the constructed ensemble prediction model, and the prediction results are presented in Figure 5.14.



```
----------------------Hypertention Prediction-----------------------
   ID  Hypertension (%)                          Unhealthy Lifestyles
0   0             93.31  Overweight, Smoking, Drinking, No Exercise
1   1             13.10
2   2             26.37
3   3             17.66
4   4              8.14
5   5             37.00
6   6             82.38                     Overweight, No Exercise
7   9             23.32
```

**Figure 5.14:** The prediction results in the simple application scenario

We introduced the operation of the proposed prediction framework using a simple application case. Then we employed SHapley Additive exPlanations

(SHAP) [188] to further analyze the contribution of high-risk population characteristics to hypertensive disorders. Specifically, SHAP (SHapley Additive exPlanations) is a game theoretical approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. The features' contribution to hypertension of those two persons (ID=0 and ID=6) with high risk are shown in 5.16.

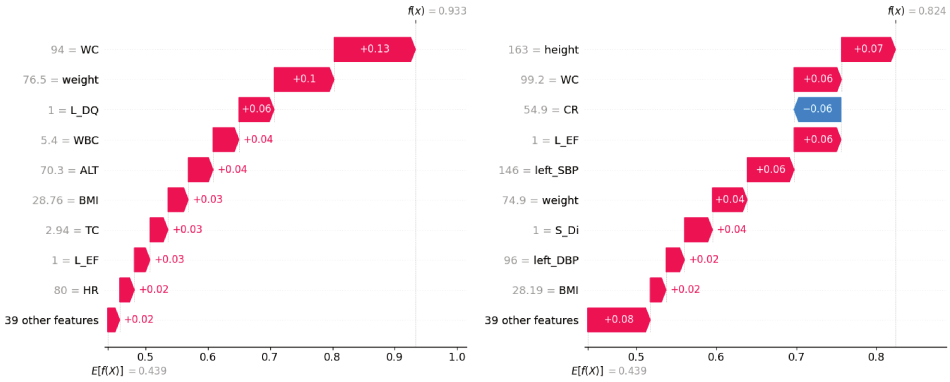



**Figure 5.15:** The features' contribution for hypertension of two persons

The above explanation shows features each contributing to pushing the model output from the base value (the average model output over the training dataset we passed) to the model output. Features pushing the prediction higher are shown in red, and those pushing the prediction lower are in blue. Specifically, for people with ID=0, the larger WC pushes up the value of the model by 0.13, while for women with ID=6, higher wc, and weight push up the predicted value of the model respectively by 0.06 and 0.04, while a normal CR pulls down the

model's prediction by a value of 0.06. SHAP can assist doctors to understand the prediction results of the model, rather than providing a black box to doctors.

Finally, in order to manage and prevent LRDs, we will also plan to create a website that predicts LRDs. We present various modules in the thesis that can assist our website to provide effective medical services.

## 5.7 Summary

This chapter primarily serves to demonstrate the three modules—missing value, feature selection, and disease prediction—that make up the proposed prediction framework. It begins by thoroughly introducing each module before analyzing a case from Nanjing, China, and using hypertension as an example for this case study. The integrated prediction model is built using a data set that only contains the important features after the missing values in the case have been processed, analyzed, and evaluated in terms of importance. Finally, the constructed model is evaluated using a range of indicators to examine its applicability to the scenario.

# Chapter 6. Conclusions and Future Work

## 6.1 Conclusions

Lifestyle-related diseases are the conclusions drawn by developed countries after conducting a large number of epidemiological investigations on chronic non-communicable diseases. One main cause of these chronic non-communicable diseases is people's unhealthy lifestyles. These diseases include obesity, hypertension, coronary heart disease, other cardiovascular diseases, stroke, and other cerebrovascular diseases, diabetes, and some malignant tumors. These diseases are difficult to cure even with modern medicine, and seriously endanger people's lives and health. Now, healthcare has been digitized and generated massive new datasets. These include electronic medical record (EMR) systems, health declaration data, radiology images, and lab results. Health service providers can propose different approaches to predictive analysis of medical diagnosis, predictive modeling of health risks, and even prescription analysis of precision medicine by combining data from different sources. Among them, disease prediction has emerged as a crucial component of any strategy for health analysis. By predicting the occurrence of diseases, it aids medical facilities in improving patient care and lowering expenditures. The development of evidence-based best practices and aiding in the identification of people at risk for lifestyle-related diseases are two

areas where disease prediction has enormous potential. This makes it possible for data to assist clinicians in staying one step ahead and offering patients proactive care before their health issues become serious.

The significant dataset noise and missing values make it challenging to use conventional machine learning methods when building LRDs prediction models utilizing medical data. Particularly, some inescapable causes, including early subject withdrawal from medical research, might quickly result in missing values in research data. Many approaches to coping with missing values have been put forth since the presence of missing values makes it more difficult to mine pertinent data. Large-scale datasets with mixed types and unbalanced features are common in the medical industry, nevertheless. Only a few approaches may be utilized for data of mixed types and unbalanced features at the same time, despite the fact that existing state-of-the-art methods can decrease imputation errors and increase the quality of missing data. In order to achieve this, we propose a novel missing value interpolation technique based on Adaptive Laplacian Weighted Random Forest (ALWRF) and SMOTE-NC oversampling technology. This method can improve Unbalanced prediction accuracy features by adaptively adjusting feature weights when building random forests.

Additionally, the algorithm's robustness will be impacted by the presence of noise. However, as noise is frequently present in medical data, a lot of studies has concentrated on how to handle it. Since some of the datasets of the analyzed lifestyle-related disorders correspond to real patients, it is difficult in practice to directly remove outliers. Combining ensemble methods with algorithm-level

techniques is an excellent strategy to minimize variance, bias, and noise. The performance of each model in the ensemble varies depending on the circumstance. In this method, an ensemble model partially addresses these shortcomings and outperforms each individual model on a combined basis. Therefore, an ensemble approach was used in our work to reduce data noise and increase the precision of lifestyle-related illness prediction. We propose a multi-objective iterative model selection (MoItMS) technique to maximize ensemble models' variety and accuracy at the same time. The proposed stacking-based multi-objective integration framework can offer useful data-driven methodologies to categorize patients for population health management, promote disease control, and support the detection of LRDs when applied to large clinical datasets.

Finally, we use a case from China to apply the proposed prediction framework. Two significant models—missing value imputation models and disease prediction models—are produced following processing by the three primary modules of missing value, feature selection, and disease prediction. The proposed prediction framework can also enhance LRDs' predicting performance for better public health prevention, according to the experimental results.

## 6.2 Future Work

### 6.2.1 Designing of LRDs Risk Prediction Website

In order to demonstrate the generalizability of the proposed approach, the study also lacks a long-term perspective on various use cases (chronic diseases

other than hypertension). Assessments in practice (multidisciplinary collaboration with clinicians) will be taken into consideration in the future within the context of this study and will require human or professional analysis. This study is going to create a website for LRD prediction in order to control and prevent LRDs. The proposed framework can assist the site in offering high-quality healthcare services. The website will include the following 7 key functions:

- User registration/login. Users need to register and log in to use the functions of the website.
- Personal information entry. Users need to enter their basic information, including name, gender, age, height, weight, blood pressure, heart rate, and other indicators.
- Disease selection. Users need to select the type of disease to be predicted, such as hypertension, diabetes, etc.
- Risk prediction. According to the information provided by the user and the type of disease selected, the website will use a predictive model to calculate the probability of the user suffering from the disease and provide corresponding suggestions.
- Health Advice. According to the information and prediction results provided by users, the website will give corresponding health advice, including diet, exercise, living habits, and so on.

- Health information. The website will regularly update health information and provide knowledge and advice on health.

The technical implementation of the LRDs risk prediction website includes four important parts.

- The front end of the website will be implemented using HTML, CSS, JavaScript, and other technologies, and adopts a responsive design to adapt to different devices and screen sizes.

- The back-end of the website will be implemented with Python language and Django framework, including user management, data management, prediction model, and other functions.

- The data of the website will be stored in a MySQL database, including user information, prediction results, health advice, etc.

- The Prediction model of the website will be implemented using the proposed forecasting framework, which can be trained according to different disease types and data provided by users to improve the accuracy of forecasting.

In order to ensure the security of user information, the website uses SSL certificates for encrypted transmission, and at the same time backs up and encrypts user data. This website aims to help users better understand their physical conditions and risks, and provide corresponding health advice and information, but it cannot replace the doctor's diagnosis and treatment. Users should treat it with caution when using it, and consult a professional in time if they have any questions doctor.

### 6.2.2 Considering Medical Data with Multiple Structures

The long-term objective of this study is to take data from multiple structures into account as this can provide more comprehensive feature information, such as fundamental knowledge, clinical examination, physiological indicators, imaging data, etc., that can be used to predict LRDs disease. The data can more accurately reflect both the disease's progression and the patient's physical state. In addition, by combining deep learning and traditional machine learning techniques, collecting feature information from various levels, and improving the model's accuracy and reliability, more sophisticated prediction models can be created using data from various structures.

# LIST OF PUBLICATIONS

The contributions proposed in this thesis have been validated by the publication of two journal papers and two conference papers as shown hereunder.

**Journal papers:**

1. Lijuan REN, Aicha SEKHARI, Haiqing ZHANG, Tao WANG, Abdelaziz BOURAS. An adaptive Laplacian weight random forest imputation for imbalance and mixed-type data, DOI: https://doi.org/10.1016/j.is.2022.102122. Information Systems, 2023, 111: 102122. [IF:3.18]

2. Lijuan REN, Haiqing ZHANG, Aicha SEKHARI, Tao WANG, Abdelaziz BOURAS. Stacking-based multi-objective ensemble framework for prediction of hypertension, DOI: https://doi.org/10.1016/j.eswa.2022.119351. Expert Systems with Applications, 2022: 119351. [IF:8.67]

3. Lijuan REN, Tao WANG, Aicha SEKHARI, Haiqing ZHANG, Abdelaziz BOURAS. A Review on Missing Values for Main Challenges and Methods. Knowledge and information systems.(Submitted).[IF:2.531]

**Conference papers:**

1. Lijuan REN, Tao WANG, Aicha SEKHARI, Haiqing ZHANG, Abdelaziz BOURAS. (2022, May). Missing Values for Classification of Machine Learning in Medical data. In 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 101-106). IEEE. DOI: 10.1109/ICAIBD 55127.2022.9820448.

2. Lijuan REN, Aicha SEKHARI, Haiqing ZHANG, Tao WANG, Abdelaziz BOURAS. Hypertension Prediction Using Optimal Random Forest and Real

Medical Data. 2022 14th 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE, 2022. [Publishing]

# References


[1] Shankar Prinja, Ruby Nimesh, Aditi Gupta, Pankaj Bahuguna, Jarnail Singh Thakur, Madhu Gupta, and Tarundeep Singh. Impact assessment and cost-effectiveness of m-health application used by community health workers for maternal, newborn and child health care services in rural uttar pradesh, india: a study protocol. *Global health action*, 9(1):31473, 2016.

[2] LJ Cortis, PR Ward, RA McKinnon, and Bogda Koczwara. Integrated care in cancer: what is it, how is it used and where are the gaps? a textual narrative literature synthesis. *European journal of cancer care*, 26(4):e12689, 2017.

[3] Junfei Chu, Xiaoxue Li, and Zhe Yuan. Emergency medical resource allocation among hospitals with non-regressive production technology: A dea-based approach. *Computers & Industrial Engineering*, 171:108491, 2022.

[4] Pijush Kanti Dutta Pramanik, Bijoy Kumar Upadhyaya, Saurabh Pal, and Tanmoy Pal. Internet of things, smart sensors, and pervasive systems: Enabling connected and pervasive healthcare. In *Healthcare data analytics and management*, pages 1–58. Elsevier, 2019.

[5] Saleem Ahmed, Kaushal Sanghvi, and Danson Yeo. Telemedicine takes centre stage during covid-19 pandemic. *BMJ Innovations*, 6(4), 2020.

[6] Zaoli Yang, Tingting Zhang, Harish Garg, and K Venkatachalam. A multi-criteria framework for addressing digitalization solutions of medical system under interval-valued t-spherical fuzzy information. *Applied Soft Computing*, page 109635, 2022.

[7] Asmat Ara Shaikh, Amala Nirmal Doss, Muthukumar Subramanian, Vipin Jain, Mohd Naved, and Md Khaja Mohiddin. Major applications of data mining in medical. *Materials Today: Proceedings*, 56:2300–2304, 2022.

[8] Qian Zhang, Anran Huang, Lianyou Shao, Peiliang Wu, Ali Asghar Heidari, Zhennao Cai, Guoxi Liang, Huiling Chen, Fahd S Alotaibi, Majdi Mafarja, et al. A machine learning framework for identifying influenza pneumonia

from bacterial pneumonia for medical decision making. *Journal of Computational Science*, page 101871, 2022.

[9] José A Castellanos-Garzón, Ernesto Costa, Juan M Corchado, et al. An evolutionary framework for machine learning applied to medical data. *Knowledge-Based Systems*, 185:104982, 2019.

[10] Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M Afshar Alam, Safdar Tanweer, and Guojun Wang. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering & Physics*, page 103825, 2022.

[11] Yann-Mickael Dalmat. Brève : La seine-et-marne, un désert médical ? *Option/Bio*, 32(643):6, 2021.

[12] Mohamed Elhoseny, K Shankar, and J Uthayakumar. Intelligent diagnostic prediction and classification system for chronic kidney disease. *Scientific reports*, 9(1):9583, 2019.

[13] M Sagner, D Katz, G Egger, L Lianov, K Schulz, H, M Braman, B Behbod, E Phillips, W Dysinger, and D and Ornish. Lifestyle medicine potential for reversing a world of chronic disease epidemics: from cell to community. *International Journal of Clinical Practice*, pages 1289–1292, 2014.

[14] Byung-Il Yeh and In Deok Kong. The advent of lifestyle medicine. *Journal of lifestyle medicine*, pages 1–8, 2013.

[15] Earl S Ford, Manuela M Bergmann, Janine Kröger, Anja Schienkiewitz, Cornelia Weikert, and Heiner Boeing. Healthy living is the best revenge: findings from the european prospective investigation into cancer and nutrition-potsdam study. *Archives of internal medicine*, 169(15):1355–1362, 2009.

[16] Alvaro Sanchez, Paola Bully, Catalina Martinez, and Gonzalo Grandes. Effectiveness of physical activity promotion interventions in primary care: a review of reviews. *Preventive medicine*, 76:S56–S67, 2015.

[17] Abdallah S Daar, Peter A Singer, Stig K Persad, Deepa Leah Pramming, David R Matthews, Robert Beaglehole, Alan Bernstein, Leszek K Borysiewicz, Stephen Colagiuri, Nirmal Ganguly, Roger I Glass, Diane T

Finegood, Jeffrey Koplan, Elizabeth G Nabel, George Sarna, Nizal Sarrafzadegan, Richard Smith, Derek Yach, and John Bell. Grand challenges in chronic non-communicable diseases. *Nature*, pages 494–496, 2007.

[18] G O'Donoghue, C Cunningham, F Murphy, C Woods, and J Aagaard-Hansen. Assessment and management of risk factors for the prevention of lifestyle-related disease: a cross-sectional survey of current activities, barriers and perceived training needs of primary care physiotherapists in the republic of ireland. *Physiotherapy*, 100(2):116–122, 2014.

[19] S Wilds, G Roglic, A Green, R Sicree, and H King. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care*, 27(5):1047–53, 2004.

[20] Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788, 2018.

[21] Shama Kakkar, Runjhun Tandon, and Nitin Tandon. The rising status of edible seeds in lifestyle related diseases: A review. *Food Chemistry*, page 134220, 2022.

[22] Ala Alwan et al. *Global status report on noncommunicable diseases 2010.* World Health Organization, 2011.

[23] Karen Morgan, Hannah Mcgee, Patrick Dicker, Ruairí Brugha, Mark Ward, Emer Shelley, Eric Van Lente, Janas Harrington, Margaret Barry, Ivan Perry, et al. Slán 2007: survey of lifestyle, attitudes and nutrition in ireland alcohol use in ireland: a profile of drinking patterns and alcohol-related harm from slán 2007. 2009.

[24] H Britt, G Miller, J Charles, Y Pan, L Valenti, J Henderson, C Bayram, J O'Halloran, and S Knox. General practice series no 19 aihw cat no gep 19. canberra: Australian institute of health and welfare; 2007. *General practice activity in Australia*, 6, 2005.

[25] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

[26] Xiao-He Hou, Lei Feng, Can Zhang, Xi-Peng Cao, Lan Tan, and Jin-Tai Yu. Models for predicting risk of dementia: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(4):373–379, 2019.

[27] Jeanne Truett, Jerome Cornfield, and William Kannel. A multivariate analysis of the risk of coronary heart disease in framingham. *Journal of chronic diseases*, 20(7):511–524, 1967.

[28] Thomas J Wang, Philimon Gona, Martin G Larson, Geoffrey H Tofler, Daniel Levy, Christopher Newton-Cheh, Paul F Jacques, Nader Rifai, Jacob Selhub, Sander J Robins, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine*, 355(25):2631–2639, 2006.

[29] Paul M Ridker, Julie E Buring, Nader Rifai, and Nancy R Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *Jama*, 297(6):611–619, 2007.

[30] Philip A Wolf, Ralph B D'Agostino, Albert J Belanger, and William B Kannel. Probability of stroke: a risk profile from the framingham study. *Stroke*, 22(3):312–318, 1991.

[31] PM Clarke, AM Gray, A Briggs, AJ Farmer, P Fenn, RJ Stevens, DR Matthews, IM Stratton, and RR Holman. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the united kingdom prospective diabetes study (ukpds) outcomes model (ukpds no. 68). *Diabetologia*, 47(10):1747–1759, 2004.

[32] Colditz GA, Atwood KA, et al. Harvard report on cancer prevention volume 4: Harvard cancer risk index. risk index working group, harvard center for cancer prevention. *Cancer Causes and Control*, 11(6):477–488, 2000.

[33] Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[34] Margaret R Spitz, Waun Ki Hong, Christopher I Amos, Xifeng Wu, Matthew B Schabath, Qiong Dong, Sanjay Shete, and Carol J Etzel. A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, 99(9):715–726, 2007.

[35] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J Khoury. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, 10(1):1–7, 2010.

[36] Vardhan Shorewala. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26:100655, 2021.

[37] Zhanlin ZHANG, Yong SUN, Xiaoqing TUO, et al. Predictive value of random forest algorithms for diabetic risk in people underwent physical examination. *Chinese General Practice*, 22(9):1021, 2019.

[38] Md Merajul Islam, Md Jahanur Rahman, Dulal Chandra Roy, Most Tawabunnahar, Rubaiyat Jahan, NAM Faisal Ahmed, and Md Maniruzzaman. Machine learning algorithm for characterizing risks of hypertension, at an early stage in bangladesh. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(3):877–884, 2021.

[39] Shahid Mohammad Ganie and Majid Bashir Malik. An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, 2:100092, 2022.

[40] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[41] V Anuja Kumari and R Chitra. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2):1797–1801, 2013.

[42] Kalliopi V Dalakleidi, Konstantia Zarkogianni, Vassilios G Karamanos, Anastasia C Thanopoulou, and Konstantina S Nikita. A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in type 2 diabetes patients. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE, 2013.

[43] Earl S Ford. Trends in predicted 10-year risk of coronary heart disease and cardiovascular disease among us adults from 1999 to 2010. *Journal of the American College of Cardiology*, 61(22):2249–2252, 2013.

[44] Hsin-Yao Wang, Shih-Cheng Chang, Wan-Ying Lin, Chun-Hsien Chen, Szu-Hsien Chiang, Kai-Yao Huang, Bo-Yu Chu, Jang-Jih Lu, and Tzong-Yi Lee. Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing. *Journal of Computational Biology*, 25(12):1347–1360, 2018.

[45] Tamara M Dugan, S Mukhopadhyay, Aaron Carroll, and Stephen Downs. Machine learning techniques for prediction of early childhood obesity. *Applied clinical informatics*, 6(03):506–520, 2015.

[46] Nongyao Nai-Arun and Rungruttikarn Moungmai. Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, 69:132–142, 2015.

[47] Todd Lingren, Vidhu Thaker, Cassandra Brady, Bahram Namjou, Stephanie Kennebeck, Jonathan Bickel, Nandan Patibandla, Yizhao Ni, Sara L Van Driest, Lixin Chen, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Applied clinical informatics*, 7(03):693–706, 2016.

[48] Daniel LaFreniere, Farhana Zulkernine, David Barber, and Ken Martin. Using machine learning to predict hypertension from a clinical dataset. In *2016 IEEE symposium series on computational intelligence (SSCI)*, pages 1–7. IEEE, 2016.

[49] Erkki Vartiainen, Tiina Laatikainen, Markku Peltonen, and Pekka Puska. Predicting coronary heart disease and stroke: the finrisk calculator. *Global heart*, 11(2):213–216, 2016.

[50] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

[51] Casimiro Aday Curbelo Montañez, Paul Fergus, Abir Hussain, Dhiya Al-Jumeily, Basma Abdulaimma, Jade Hind, and Naeem Radi. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2743–2750. IEEE, 2017.

[52] Kunal Rajput, Girija Chetty, and Rachel Davey. Obesity and co-morbidity detection in clinical text using deep learning and machine learning techniques. In *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pages 51–56. IEEE, 2018.

[53] Chengyin Ye, Tianyun Fu, Shiying Hao, Yan Zhang, Oliver Wang, Bo Jin, Minjie Xia, Modi Liu, Xin Zhou, Qian Wu, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res*, 20(1):e22, 2018.

[54] Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):1–7, 2018.

[55] Renuka Patnaik, Mahesh Chandran, Seung-Cheol Lee, Anurag Gupta, Chansoo Kim, and Changsoo Kim. Predicting the occurrence of essential hypertension using annual health records. In *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–5. IEEE, 2018.

[56] Fernando López-Martínez, Aron Schwarcz, Edward Rolando Núñez-Valdez, and Vicente Garcia-Diaz. Machine learning classification analysis for a hypertensive population as a function of several risk factors. *Expert Systems with Applications*, 110:206–215, 2018.

[57] Valery Effoe, Alain Bertoni, Fengxia Yan, Nchang Taka, Obiora Egbuche, Titilope Olanipekun, Demilade Adedinsewo, Ervin Fox, and Herman Taylor. Carotid intima-media thickness versus coronary artery calcium score in pre-

dicting incident coronary heart disease and stroke among african americans: The jackson heart study. *Journal of the American College of Cardiology*, 71(11S):A1872–A1872, 2018.

[58] Isaac Machorro-Cano, Giner Alor-Hernández, Mario Andrés Paredes-Valverde, Uriel Ramos-Deonati, José Luis Sánchez-Cervantes, and Lisbeth Rodríguez-Mazahua. Pisiot: a machine learning and iot-based smart health platform for overweight and obesity control. *Applied Sciences*, 9(15):3037, 2019.

[59] Othmane Daanouni, Bouchaib Cherradi, and Amal Tmiri. Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In *Proceedings of the 4th International Conference on Smart City Applications*, pages 1–6, 2019.

[60] Ravinder Ahuja, Subhash C Sharma, and Maaruf Ali. A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, pages 2516–0281, 2019.

[61] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, and Mirsat Yesiltepe. A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International informatics and software engineering conference (UBMYK)*, pages 1–4. IEEE, 2019.

[62] Fernando López-Martínez, Edward Rolando Núñez-Valdez, Rubén González Crespo, and Vicente García-Díaz. An artificial neural network approach for predicting hypertension using nhanes data. *Scientific Reports*, 10(1):1–14, 2020.

[63] Hendrana Tjahjadi and Kalamullah Ramli. Noninvasive blood pressure classification based on photoplethysmography using k-nearest neighbors algorithm: a feasibility study. *Information*, 11(2):93, 2020.

[64] Kezban Alpan and Galip Savaş İlgi. Classification of diabetes dataset with data mining techniques by using weka approach. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–7. IEEE, 2020.

[65] Motiur Rahman, Dilshad Islam, Rokeya Jahan Mukti, and Indrajit Saha. A deep learning approach based on convolutional lstm for detecting diabetes. *Computational biology and chemistry*, 88:107329, 2020.

[66] Shahan Ali Memon, Saquib Razak, and Ingmar Weber. Lifestyle disease surveillance using population search behavior: Feasibility study. *Journal of medical Internet research*, 22(1):e13347, 2020.

[67] Balbir Singh and Hissam Tawfik. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, pages 523–535. Springer, 2020.

[68] Arvind Kumar Shukla. Patient diabetes forecasting based on machine learning approach. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2019*, pages 1017–1027. Springer, 2020.

[69] MM Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer vision and machine intelligence in medical image analysis*, pages 113–125. Springer, 2020.

[70] Mohamed Abdel-Basset, Rehab Mohamed, Abd El-Nasser H Zaied, Abduallah Gamal, and Florentin Smarandache. Solving the supply chain problem using the best-worst method based on a novel plithogenic model. In *Optimization theory based on neutrosophic and plithogenic sets*, pages 1–19. Elsevier, 2020.

[71] Ali Aminian, Alexander Zajichek, David E Arterburn, Kathy E Wolski, Stacy A Brethauer, Philip R Schauer, Steven E Nissen, and Michael W Kattan. Predicting 10-year risk of end-organ complications of type 2 diabetes with and without metabolic surgery: a machine learning approach. *Diabetes Care*, 43(4):852–859, 2020.

[72] Maria Athanasiou, Konstantina Sfrintzeri, Konstantia Zarkogianni, Anastasia C Thanopoulou, and Konstantina S Nikita. An explainable xgboost–based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus. In *2020 IEEE 20th International*

*Conference on Bioinformatics and Bioengineering (BIBE)*, pages 859–864. IEEE, 2020.

[73] Mehrdad Rezaee, Igor Putrenko, Arsia Takeh, Andrea Ganna, and Erik Ingelsson. Development and validation of risk prediction models for multiple cardiovascular diseases and type 2 diabetes. *PLoS one*, 15(7):e0235758, 2020.

[74] Akkem Yaganteeswarudu. Multi disease prediction model by using machine learning and flask api. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1242–1246. IEEE, 2020.

[75] Luís Chaves and Gonçalo Marques. Data mining techniques for early diagnosis of diabetes: a comparative study. *Applied Sciences*, 11(5):2218, 2021.

[76] Ke Wang, Jing Tian, Chu Zheng, Hong Yang, Jia Ren, Yanling Liu, Qinghua Han, and Yanbo Zhang. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and shap. *Computers in Biology and Medicine*, 137:104813, 2021.

[77] Md. Merajul Islam, Md. Jahanur Rahman, Dulal Chandra Roy, Most. Tawabunnahar, Rubaiyat Jahan, N.A.M.Faisal Ahmed, and Md. Maniruzzaman. Machine learning algorithm for characterizing risks of hypertension, at an early stage in bangladesh. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, 15(3):877–884, 2021.

[78] Liying Li, Ziqiong Wang, Muxin Zhang, Haiyan Ruan, Linxia Zhou, Xin Wei, Ye Zhu, Jiafu Wei, and Sen He. New risk score model for identifying individuals at risk for diabetes in southwest china. *Preventive Medicine Reports*, 24:101618, 2021.

[79] Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, and Md. Tarek Habib. A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, 2:100053, 2021.

[80] Junaid Rashid, Saba Batool, Jungeun Kim, Muhammad Wasif Nisar, Amir Hussain, Sapna Juneja, and Riti Kushwaha. An augmented artificial intel-

ligence approach for chronic diseases prediction. *Frontiers in Public Health*, 10:559, 2022.

[81] Aditya Gupta and Amritpal Singh. An optimal multi-disease prediction framework using hybrid machine learning techniques: 10.48129/kjs. splml. 19321. *Kuwait Journal of Science*, 2022.

[82] Jingjing Yan, Jing Tian, Hong Yang, Gangfei Han, Yanling Liu, Hangzhi He, Qinghua Han, and Yanbo Zhang. A clinical decision support system for predicting coronary artery stenosis in patients with suspected coronary heart disease. *Computers in Biology and Medicine*, 151:106300, 2022.

[83] Guangwei Li, Ping Zhang, Jinping Wang, Edward W Gregg, Wenying Yang, Qiuhong Gong, Hui Li, Hongliang Li, Yayun Jiang, Yali An, et al. The long-term effect of lifestyle interventions to prevent diabetes in the china da qing diabetes prevention study: a 20-year follow-up study. *The Lancet*, 371(9626):1783–1789, 2008.

[84] Fan Cheng and Yiwei Yin. Application of computer data analysis technology in the development of a physical education examination platform. *International Journal of Emerging Technologies in Learning*, 14(6), 2019.

[85] Xiao-Ling Wang, Jun Liu, Zi-Qi Li, and Zhi-Lin Luan. Application of physical examination data on health analysis and intelligent diagnosis. *BioMed Research International*, 2021, 2021.

[86] Hui Yang, Yamei Luo, Xiaolei Ren, Ming Wu, Xiaolin He, Bowen Peng, Kejun Deng, Dan Yan, Hua Tang, and Hao Lin. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Information Fusion*, 75:140–149, 2021.

[87] Mariko Kawasoe, Shin Kawasoe, Takuro Kubozono, Satoko Ojima, Takeko Kawabata, Yoshiyuki Ikeda, Naoya Oketani, Hironori Miyahara, Koichi Tokushige, Masaaki Miyata, et al. Development of a risk prediction score for hypertension incidence using japanese health checkup data. *Hypertension Research*, 45(4):730–740, 2022.

[88] Xin Qian, Yu Li, Xianghui Zhang, Heng Guo, Jia He, Xinping Wang, Yizhong Yan, Jiaolong Ma, Rulin Ma, and Shuxia Guo. A cardiovascular

disease prediction model based on routine physical examination indicators using machine learning methods: A cohort study. *Frontiers in cardiovascular medicine*, 9, 2022.

[89] Mohammad Shafenoor Amin, Yin Kia Chiam, and Kasturi Dewi Varathan. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, 36:82–93, 2019.

[90] Zhaozhao Xu, Derong Shen, Tiezheng Nie, Yue Kou, Nan Yin, and Xi Han. A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data. *Information Sciences*, 572:574–589, 2021.

[91] Bhavisha Suthar, Hemant Patel, and Ankur Goswami. A Survey: Classification of Imputation Methods in Data Mining. *International Journal of Emerging Technology and Advanced Engineering*, 2(1):309–312, 2012.

[92] Rima Houari, Ahcène Bounceur, A. Kamel Tari, and M. Tahar Kecha. Handling missing data problems with sampling methods. *Proceedings - 2014 International Conference on Advanced Networking Distributed Systems and Applications, INDS 2014*, pages 99–104, 2014.

[93] Tlamelo Emmanuel. A Survey On Missing Data in Machine Learning. *Research Square*, 8(1):1–37, 2021.

[94] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

[95] Roderick J.A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics, 2014.

[96] Schafer Joseph L. and Graham John W. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[97] Tero Aittokallio. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2):253–264, 12 2009.

[98] Jason Van Hulse and Taghi M. Khoshgoftaar. A comprehensive empirical evaluation of missing value imputation in noisy software measurement data.

*Journal of Systems and Software*, 81(5):691–708, 2008. Software Process and Product Measurement.

[99] Iris Eekhout, Henrica C.W. de Vet, Jos W.R. Twisk, Jaap P.L. Brand, Michiel R. de Boer, and Martijn W. Heymans. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, 67(3):335–342, 2014.

[100] Trond Hellem Bø, Bjarte Dysvik, and Inge Jonassen. Lsimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic acids research*, 32(3):e34–e34, 2004.

[101] Pedro J. García-Laencina, José Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[102] Tobias Rockel, Dieter William Joenssen, and Udo Bankhofer. Decision Trees for the Imputation of Categorical Data. *Kit Scientific Publishing*, 2(1):1–15, 2017.

[103] Xianping Du, Hongyi Xu, and Feng Zhu. A data mining method for structure design with uncertainty in design variables. *Computers and Structures*, 244:106457, 2021.

[104] Kancherla Jonah Nishanth, Vadlamani Ravi, Narravula Ankaiah, and Indranil Bose. Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. *Expert Systems with Applications*, 39(12):10583–10589, 2012.

[105] Bahareh Fallah, Kelvin Tsun Wai Ng, Hoang Lan Vu, and Farshid Torabi. Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation. *Waste Management*, 116:66–78, 2020.

[106] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, and H. Lähdesmäki. Self-organization and missing values in som and gtm. *Neurocomputing*, 147:60–70, 2015. Advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012).

[107] Khaled Mohamad Almustafa. Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, 24:100631, 2021.

[108] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[109] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[110] Yufei Xia, Chuanzhe Liu, YuYing Li, and Nana Liu. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78:225–241, 2017.

[111] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[112] Shahla Faisal and Gerhard Tutz. Imputation methods for high-dimensional mixed-type datasets by nearest neighbors. *Computers in Biology and Medicine*, page 104577, 2021.

[113] Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowledge-Based Systems*, 53:51–65, 2013.

[114] Sanaz Nikfalazar, Chung Hsing Yeh, Susan Bedingfield, and Hadi A. Khorshidi. Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, 62(6):2419–2437, 2020.

[115] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[116] Vincent Cabeli, Louis Verny, Nadir Sella, Guido Uguzzoni, Marc Verny, and Hervé Isambert. Learning clinical networks from medical records based

on information estimates in mixed-type data. *PLoS computational biology*, 16(5):e1007866, 2020.

[117] Der-Chiang Li, Chiao-Wen Liu, and Susan C Hu. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, 40(5):509–518, 2010.

[118] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210, 2004.

[119] José A Sáez, Mikel Galar, Julián Luengo, and Francisco Herrera. Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition. *Knowledge and information systems*, 38(1):179–206, 2014.

[120] Mohammed Gollapalli, Aisha Alansari, Heba Alkhorasani, Meelaf Alsubaii, Rasha Sakloua, Reem Alzahrani, Mohammed Al-Hariri, Maiadah Alfares, Dania AlKhafaji, Reem Al Argan, et al. A novel stacking ensemble for detecting three types of diabetes mellitus using a saudi arabian dataset: prediabetes, t1dm, and t2dm. *Computers in Biology and Medicine*, 147:105757, 2022.

[121] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[122] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[123] Taghi M Khoshgoftaar and Pierre Rebours. Improving software quality prediction by noise filtering techniques. *Journal of Computer Science and Technology*, 22(3):387–396, 2007.

[124] Sabina Bijlsma, Ivana Bobeldijk, Elwin R Verheij, Raymond Ramaker, Sunil Kochhar, Ian A Macdonald, Ben Van Ommen, and Age K Smilde. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical chemistry*, 78(2):567–574, 2006.

[125] Linqi Zhu, Xueqing Zhou, and Chaomo Zhang. Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. *Artificial Intelligence in Geosciences*, 2:76–81, 2021.

[126] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[127] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[128] Runhai Feng, Dario Grana, and Niels Balling. Imputation of missing well log data by random forest and its uncertainty analysis. *Computers Geosciences*, 152:104763, 2021.

[129] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[130] Dhammika Amaratunga, Javier Cabrera, and Yung-Seop Lee. Enriched random forests. *Bioinformatics*, 24(18):2010–2014, 07 2008.

[131] Yunming Ye, Qingyao Wu, Joshua Zhexue Huang, Michael K Ng, and Xutao Li. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, 46(3):769–787, 2013.

[132] Jianheng Liang and Dong Huang. Laplacian-weighted random forest for high-dimensional data classification. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 748–753, 2019.

[133] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

[134] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[135] Steven L Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, 1994.

[136] L Breiman, JH Friedman, RA Olshen, and CJ Stone. Cart. *Classification and Regression Trees, Wadsworth and Brooks/Cole, Monterey, CA*, 1984.

[137] Georgios Douzas, Fernando Bacao, and Felix Last. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, 465:1–20, 2018.

[138] Hong Zhao and Xiangju Li. A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism. *Information Sciences*, 378:303–316, 2017.

[139] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

[140] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.

[141] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[142] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011.

[143] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[144] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[145] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[146] H Golino. Women's dataset from the'predicting increased blood pressure using machine learning, figshare, 2013.

[147] H Golino. Men's dataset from the'predicting increased blood pressure using machine learning, figshare, 2013.

[148] Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the

onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

[149] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[150] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[151] A. Frank and A. Asuncion. Uci machine learning repository. 2010.

[152] Md Geaur Rahman and Md Zahidul Islam. Data quality improvement by imputation of missing values. *International Conference on Computer Science and Information Technology*, pages 82–88, 2013.

[153] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Boosting and additive trees. In *The elements of statistical learning*, pages 337–387. Springer, 2009.

[154] Miroslav Kubat. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, 13(4):409–412, 1999.

[155] Mark H Licht. Multiple regression and correlation. 1995.

[156] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714, 2010.

[157] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[158] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[159] Palanisamy Shunmugapriya and S Kanmani. Optimization of stacking ensemble configurations through artificial bee colony algorithm. *Swarm and Evolutionary Computation*, 12:24–32, 2013.

[160] Yijun Chen, Man-Leung Wong, and Haibing Li. Applying ant colony optimization to configuring stacking ensembles for data mining. *Expert systems with applications*, 41(6):2688–2702, 2014.

[161] Renata Furtuna, Silvia Curteanu, and Florin Leon. Multi-objective optimization of a stacked neural network using an evolutionary hyper-heuristic. *Applied Soft Computing*, 12(1):133–144, 2012.

[162] Shasha Mao, Jia-Wei Chen, Licheng Jiao, Shuiping Gou, and Rongfang Wang. Maximizing diversity by transformed ensemble learning. *Applied Soft Computing*, 82:105580, 2019.

[163] Huanhuan Chen and Xin Yao. Multiobjective neural network ensembles based on regularized negative correlation learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(12):1738–1751, 2010.

[164] Fei Li, Li Zhang, Bin Chen, Dianzhu Gao, Yijun Cheng, Xiaoyong Zhang, Yingze Yang, Kai Gao, and Zhiwu Huang. An optimal stacking ensemble for remaining useful life estimation of systems under multi-operating conditions. *IEEE Access*, 8:31854–31868, 2020.

[165] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[166] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[167] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 12, 1999.

[168] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[169] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[170] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[171] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[172] Yijun Chen and Man Leung Wong. An ant colony optimization approach for stacking ensemble. In *2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC)*, pages 146–151. IEEE, 2010.

[173] Francisco Javier Ordóñez, Agapito Ledezma, and Araceli Sanchis. Genetic approach for optimizing ensembles of classifiers. In *FLAIRS conference*, pages 89–94, 2008.

[174] WMAW Ahmad, MABA Nawi, N Aleng, N Halim, Mustafa Mamat, M Hamzah, and Zalila Ali. Association of hypertension with risk factors using logistic regression. *Applied Mathematical Sciences*, 8(52):2563–2572, 2014.

[175] Latifa A AlKaabi, Lina S Ahmed, Maryam F Al Attiyah, and Manar E Abdel-Rahman. Predicting hypertension using machine learning: Findings from qatar biobank study. *Plos one*, 15(10):e0240370, 2020.

[176] Daniel H Katz, Rahul C Deo, Frank G Aguilar, Senthil Selvaraj, Eva E Martinez, Lauren Beussink-Nelson, Kwang-Youn A Kim, Jie Peng, Marguerite R Irvin, Hemant Tiwari, et al. Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. *Journal of cardiovascular translational research*, 10(3):275–284, 2017.

[177] Shuqiong Huang, Yihua Xu, Li Yue, Sheng Wei, Li Liu, Xiumin Gan, Shuihong Zhou, and Shaofa Nie. Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a chinese area. *Hypertension Research*, 33(7):722–726, 2010.

[178] Hudson Fernandes Golino, Liliany Souza de Brito Amaral, Stenio Fernando Pimentel Duarte, Cristiano Mauro Assis Gomes, Telma de Jesus

Soares, Luciana Araujo dos Reis, and Joselito Santos. Predicting increased blood pressure using machine learning. *Journal of obesity*, 2014, 2014.

[179] Enid Wai-Yung Kwong, Hao Wu, and Grantham Kwok-Hung Pang. A prediction model of blood pressure for telemedicine. *Health informatics journal*, 24(3):227–244, 2018.

[180] Chih-Ta Yen, Sheng-Nan Chang, and Cheng-Hong Liao. Deep learning algorithm evaluation of hypertension classification in less photoplethysmography signals conditions. *Measurement and Control*, 54(3-4):439–445, 2021.

[181] Jaypal Singh Rajput, Manish Sharma, and U Rajendra Acharya. Hypertension diagnosis index for discrimination of high-risk hypertension ecg signals using optimal orthogonal wavelet filter bank. *International journal of environmental research and public health*, 16(21):4068, 2019.

[182] Cut Fiarni, Evasaria M Sipayung, and Siti Maemunah. Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia computer science*, 161:449–457, 2019.

[183] Dhammika Amaratunga, Javier Cabrera, Davit Sargsyan, John B Kostis, Stavros Zinonos, and William J Kostis. Uses and opportunities for machine learning in hypertension research. *International Journal of Cardiology Hypertension*, 5:100027, 2020.

[184] Jidapa Kraisangka, Lisa Carey Lohmueller, Manreet Kaur Kanwar, Carol Zhao, MJ Druzdzel, James Francis Antaki, Marc A Simon, and Raymond L Benza. Derivation of a bayesian network model from an existing risk score calculator for pulmonary arterial hypertension. *The Journal of Heart and Lung Transplantation*, 38(4):S487–S488, 2019.

[185] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122:56–69, 2020.

[186] Namrata Singh and Pradeep Singh. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, 40(1):1–22, 2020.

[187] Cox D Bergstra J, Yamins D. *Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures*. 2013.

[188] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.