Master’s Thesis of Sport Science

Identification of Differentially Methylated Genes Associated with Sarcopenia Using Machine Learning : the Korean Genome and Epidemiology Study (KoGES)

머신러닝 기법을 활용한 근감소증 DNA 메틸화 분석 :한국인유전체역학조사사업 데이터 활용

August 2023

Graduate School of Physical Education
Seoul National University
Sport Science Major

Seohyun Ahn

Identification of Differentially Methylated Genes Associated with Sarcopenia Using Machine Learning : the Korean Genome and Epidemiology Study (KoGES)

Wook Song, Ph.D.

Submitting a master's thesis of
Sport Science

August 2023

Graduate School of Physical Education
Seoul National University
Sport Science Major

Seohyun Ahn

Confirming the master's thesis written by
Seohyun Ahn
August 2023

Chair Yongho Lee (Seal)
Vice Chair Hyo Youl Moon (Seal)
Examiner Wook Song (Seal)

Abstract

With the aging population on the rise, there is a growing emphasis on geriatric illnesses, with sarcopenia being a notable focus. Comparable to physical aging, sarcopenia manifests as a decline in muscle mass, strength, and physical performance as one ages. This condition is linked to heightened mortality rates, osteoporosis, fractures, and other ailments. While once perceived as a natural aspect of getting older, recent times have officially recognized sarcopenia as a distinct medical condition. Recently, the World Health Organization (WHO) listed sarcopenia as an official disease code, and Korea has also assigned a disease code (M62.5) to sarcopenia.

As individuals age, the diminishing muscle mass, strength, and physical function elevate the risk of falls and susceptibility to secondary health issues. The pace of this decline varies among individuals and stems from a blend of fixed genetic components and environmental influences. Epigenetic inquiries are imperative to decipher the intricate interplay between genetic and environmental elements, the epigenetic factors associated with sarcopenia are still not well understood.

Epigenetics is the phenomenon of influencing gene expression without changing the DNA sequence, which is known to be influenced by environmental factors encountered throughout life. The primary mechanisms within epigenetics involve DNA methylation and histone modifications, with DNA methylation emerging as a pivotal player in the processes of biological aging and the emergence of chronic ailments. The notion of link between sarcopenia and DNA methylation has been introduced, driven by observations that alterations in muscle mass and function, a consequence of aging and diverse diseases, coincide with DNA methylation shifts. Therefore, this study purpose to unveil potential sarcopenia biomarkers within the Korean population, utilizing data sourced from the Korean Genome Epidemiology Study (KoGES), and subsequently, to generate a diagnostic and predictive model for sarcopenia using machine learning algorithms, leveraging the genome as a foundation.

Using data from the Korea Genome and Epidemiology Study (KoGES) from 2004 to 2013. A total of 110 participants (82 male and 28 female) were included to investigate the association of the identified differentially methylated DNA probes with the occurrence of sarcopenia. Participants were categorized according to two variables: muscle mass (appendicular skeletal muscle index; ASMI) and muscle strength (handgrip) according to the criteria of Asian sarcopenia. The sarcopenia groups were determined by dividing lower and upper quantile of the total data. The differentially methylated DNA probes data were assayed with the Infinium Methylation Epic Beadchip from Infinium, and after appropriate data processing steps, including normalization and correction for DNA methylation batch effects, a total of more than 740,000 markers within genes were obtained. Differentially methylated DNA probes were subsequently analyzed using criteria of $|\mathrm{logFC}| > 0.15$ and $\mathrm{FDR\ adjust}\ p - \mathrm{value} < 0.05$. In the male group, 99 hypermethylation and 67 hypomethylation were found, but in the female group, the threshold was not met, and the differential methylation analysis could not be performed. Hence, the data from the female group was excluded due to the absence of significant results concerning differential methylation.

To identify key biomarkers, a total of 166 differential methylation probes were analyzed. To ensure that the variable data were normally distributed, 134 variables were removed using Pearson correlation. Recursive feature elimination cross-validation (RFECV) was then used to select significant variables. Finally, a total of 10 probes with significant associations were identified. Using the 10 probes, built a diagnostic model for sarcopenia using a majority voting ensemble that combines the predictions of multiple models. This ensemble technique was used because it is a technique that can be used to improve model performance and can achieve better performance than a single model. The train and test data sets were split 7:3 for analysis. The train dataset was trained using four algorithms: decision tree, random forest, logistic regression, K-Nearest Neighbors, and Naïve Bayes, and the predictions of the individual models used were combined to derive the majority voting value. Finally, the diagnostic model was evaluated using the test data set, and the diagnostic performance was evaluated using the area under the curve (AUC) value. The constructed model showed high accuracy (96%) in identifying the genetic phenotype of sarcopenia in Koreans.

In addition, three of the 10 selected probes, TCF12, RYR2, and ZNF415 were identified as potential biomarkers of sarcopenia. The

TCF12 gene has a significant impact on muscle development and regeneration, while RYR2 is known for its role in heart muscle and may also affect skeletal muscle through its interaction with RYR1 receptors. Regarding ZNF415, although it has not been extensively studied compared to other genes, it participates in various cellular processes, including gene and transcriptional regulation. Previous research has linked ZNF415 to sarcopenia, suggesting its potential association with the condition.

Furthermore, given that aging is linked to a reduction in muscle mass, strength, and overall bodily functions, examined their correlations with aging using the 10 identified probes. Among these, RYR2 exhibited the most pronounced negative correlation with age (r = -0.64), highlighting that sarcopenia is not exclusively limited to the elderly; it can also manifest in middle-aged individuals. This underscores the significance of addressing muscle aging during middle age to potentially forestall or alleviate the onset of sarcopenia.

Using machine learning techniques, potential genetic biomarkers of sarcopenia were identified, and an early prediction model with high diagnostic performance for sarcopenia was successfully developed. Despite the enhanced predictive capability for sarcopenia, conducting comprehensive cellular and molecular biological validation is crucial to elucidate the connection between the identified methylation candidate molecules and sarcopenia. The identified potential genetic biomarkers of sarcopenia offer valuable insights into the underlying mechanisms of sarcopenia risk in Koreans and serve as a valuable reference for targeted interventions.

Keywords: Sarcopenia; Differentially methylated DNA probes; Epigenome; Machine learning; Korean Genome Epidemiology Study (KoGES)

Student Number: 2021-28886

Table of Contents

List of Figure

List of Table

Definitions and Abbreviations

ASM	Appendicular skeletal muscle
ASMI	Appendicular skeletal muscle index
AUC	Area under the curve
BMI	Body mass index
CpG	A site in the DNA sequence where a cytosin (C) is adjacent to aguanine (G) base, connecter by a phosphodiester bond
DNA	Deoxyribonucleic acid
DMPs	Differentially methylated probes
dmCpGs	Differentially methylated CpG sites
DT	Decision tree
FDR adjusted *p*-value	adjusted False discovery rate
FN	False negative
FP	False positive
FPR	False positive rate
KNN	K-Nearest Neighbors
KoGES	Korean Genome Epidemiology Study
LogFC	Log fold change
NB	Naïve Bayes
RFE	Recursive Feature Elimination
RFECV	Recursive Feature Elimination with cross validation
RF	Random Forest
RNA	Ribonucleic acid
ROC	Receiver operating characteristic curve
SVM	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate

Chapter 1. Introduction

1.1. Significance of Research

Sarcopenia is a muscle disease characterized by the progressive muscle mass and strength decline, is commonly associated with aging (Chen et al., 2014). While aging can be a major factor in causing sarcopenia, it is often secondary to degenerative diseases, in addition to diseases of the muscles themselves (Alfonso J. Cruz-Jentoft et al., 2010). Despite extensive research on sarcopenia, the epigenetic factors associated with sarcopenia are still not well understood.

Epigenetics refers to the process of regulating gene expression without altering the DNA sequence and is influenced by various environmental factors encountered throughout our lifetime. DNA methylation and histone modification are the main mechanisms of epigenetics, with DNA methylation being closely associated with biological aging and the development of chronic illnesses. Sarcopenia, characterized by changes in muscle quantity and quality due to aging or other diseases, is also linked to DNA methylation. Muscle, being essential for metabolite storage and conversion, can influence DNA methylation through alterations in muscle mass or metabolite levels. For example, inflammatory cytokines, like tumor necrosis factor-alpha and interleukin-, contribute to muscle loss and impaired regeneration during aging (Kwak et al., 2018). Besides, elevated serum levels of certain markers, such as including interleukin-6, secreted protein acidic and rich in cysteine, and macrophage migration inhibitory factor, are observed in individuals with sarcopenia, while insulin-like growth factor 1 levels are lower. Thus, combination of these serum levels could potentially serve as biomarkers for sarcopenia, and studying DNA methylation in blood may provide insights into the systemic factors involved in this condition (Kwak et al., 2018). If there is an overlap in DNA methylation patterns between blood and muscle biopsy samples, easily accessible blood samples could be used for further research and potential biomarker applications (He et al., 2019).

DNA methylation is a significant epigenetic mechanism involving chemical modifications to the DNA molecule (Moore et al., 2013). It plays a dynamic role in regulating gene expression during development and differentiation (Jones et al., 2015). Methylation occurring within intragenic regions, especially CpG islands, can have various functional impacts. For instance, methylation within actively transcribed genes can hinder gene expression by reducing the efficiency of RNA polymerase II elongation (Lorincz et al., 2004). Intragenic DNA methylation also serves to prevent spurious transcriptions and can be regulated by other CpG islands acting as transcription initiators (Jeziorska et al., 2017). Additionally, CpG shores, located near CpG islands, can influence gene expression through methylation (Rao et al., 2013). Various factors, including age, lifestyle, nutrition, and environmental elements like air pollution, can influence patterns of DNA methylation (Barrès et al., 2012; Martin & Fry, 2018; Nitert et al., 2012; Santos et al., 2019; Seaborne et al., 2018).

In the medical field, the identification of potential disease biomarkers based on epigenetics has gained significant importance. These biomarkers are utilized to develop predictive risk models that aid in diagnosing and predicting diseases in individuals (Moons et al., 2012). Prediction models can serve either diagnostic or prognostic purposes, aiming to provide objective measures to support physicians in clinical decision-making (Hendriksen et al., 2013). By accurately identifying individuals at risk of disease development, physicians can implement early prevention strategies and provide personalized, targeted care to those most vulnerable. Predictive modeling aims to estimate the probability of an event outcome rather than establishing causality between features (predictors) and an outcome (Waljee et al., 2014). The process involves two phases: development and application. During the development or training phase, a model is chosen and optimized using a representative dataset to predict the desired outcome. The optimized model is then applied to make predictions on new data. Machine learning approaches are employed for predictive modeling, where algorithms recognize relationships within a subset of training data and assess the model's predictive performance on a separate test dataset. The focus is on identifying relationships within the data, with less emphasis on understanding these relationships. Machine learning, compared to traditional statistical methods, excels in addressing complex and non-linear relationships present in real-world problems, making it a valuable tool across various fields.

The main objective of this research is to explore the variations in DNA probe methylation between individuals with sarcopenia and those non- sarcopenia in the KoGES (Korean Genome Epidemiology Study) using advanced machine learning techniques. By identifying specific DNA methylation patterns associated with different gene types, this study seeks to develop predictive and diagnostic algorithms capable of characterizing sarcopenia in the Korean population based on its epigenetic profiles.

1.2. Purpose of Research

The purpose of this thesis is to develop a diagnostic model for sarcopenia in Koreans using machine learning. In order to accomplish that, will accomplish the following objectives:

(1) To identify and compare sarcopenia and non-sarcopenia to identify distinct DNA methylation patterns in sarcopenia;
(2) To select significant probe features among the differentiated DNA methylation probes using feature selection method;
(3) To build a machine learning algorithm model that can improve the classification performance using the important probe.

The identified methylation probes will provide valuable insights into the underlying mechanisms of sarcopenia risk in Koreans and provide potential references for targeted interventions.

1.3. Hypothesis of Research

The study hypotheses are as follows:

(1) Patients with sarcopenia will demonstrate distinct DNA methylation patterns in comparison to patients without sarcopenia.

(2) After identifying the differentially methylated DNA probes, a feature selection process will be conducted to extract highly relevant features associated with sarcopenia. The resulting model is expected to outperform.

(3) The selected DNA probes will be examined for their association with age, and it is anticipated that some probes will exhibit higher DNA expression in older individuals compared to middle-aged individuals.

Chapter 2. Background

2.1. Sarcopenia

2.1.1. Definition of sarcopenia

Sarcopenia is a condition characterized by the progressive loss of skeletal muscle mass, often accompanied by a decline in muscle strength and performance (Chen et al., 2014; L. K. Chen et al., 2020; Evans, 1995). The term "sarcopenia" originates from the Greek words "sarx," meaning flesh, and "penia," meaning loss, which translates to "loss of flesh" (Alfonso J Cruz-Jentoft et al., 2010; Evans, 1995). It predominantly occurs in older individuals and is considered a natural consequence of the aging process in many cases (Evans, 1995; Massimino et al., 2021). Depending on the diagnostic criteria used, the prevalence of sarcopenia globally ranges from 10% to 40% of the population (Massimino et al., 2021; Mesinovic et al., 2019). Around the age of 50, the annual rate of muscle mass loss is estimated to be between 1& and 2%, with a concurrent 1.5% annual decrease in muscle strength between 50 and 60 years of age. Some studies have even reported reductions in muscle mass and muscle function beginning around the ages of 30 or 40 (Tan, 2020; von Haehling et al., 2010). As people reach the age of 60 and beyond, muscle strength tends to decline at a rate of 3% per year (Tan, 2020; von Haehling et al., 2010). Furthermore, from the age of 70 onwards, muscle mass is estimated to decrease by around 15-25% every ten years (Izzo et al., 2021). To sum up, sarcopenia can be classified into two types: primary and secondary sarcopenia. Primary sarcopenia refers to the natural and persistent loss of muscle mass that occurs with aging (Santilli et al., 2014). Secondary sarcopenia, on the other hand, is influenced by external factors such as lifestyle, physical activity, diet, comorbid diseases, or other underlying conditions (Santilli et al., 2014). Both of these factors contribute to the expression of sarcopenia, which affects DNA methylation.

2.1.2. Diagnosis of sarcopenia

Throughout the years, various definitions and guidelines have been put forward to diagnose sarcopenia, but a unified and comprehensive set of diagnostic criteria has not been established. The assessment of three key variable are typically employed to determine the presence of sarcopenia: (1) measurement of muscle mass, (2) evaluation of muscle strength, and (3) assessment of muscle function or performance (Alfonso J Cruz-Jentoft et al., 2010; Santilli et al., 2014). In 2019, the Asian Working Group for Sarcopenia (AWGS) introduced updated guidelines by retaining the core framework while making certain modifications to the cutoff values (L.-K. Chen et al., 2020). They also provided specific recommendations for different measurement methods. According to these guidelines, Asian sarcopenia is defined as the condition when both appendicular skeletal muscle mass and muscle strength fall below baseline values, or when appendicular skeletal muscle mass and physical performance are below baseline values (Kim & Won, 2020). The Asian Working Group for Sarcopenia (AWGS) recommends cutoff values for muscle mass measurements of 7.0 kg/m^2 for male and 5.7 kg/m^2 or female using bioimpedance analysis, along with a handgrip strength cutoff of <26 kg for male and <18 kg for female.

In this study, appendicular skeletal muscle mass was measured using bioelectrical impedance analysis (BIA), and the following equation (Equation 1) derived from (Xu et al., 2011) was used to calculate limb skeletal muscle mass:

$$\begin{aligned} &\text{Appendicular skeletal muscle (ASM)} \\ &\qquad = 0.193 * \text{body weight} + 0.107 * \text{height} - 4.157 * \text{gender} \\ &\qquad - 0.037 * \text{age} - 2.631 \end{aligned}$$

weight in kg, height in m, age in year, gender:1 for men and 2 for women

Equation 1. Appendicular skeletal muscle mass calculator formula

The appendicular skeletal muscle mass index (ASMI) was calculated using the measured appendicular skeletal muscle mass (ASM). Appendicular skeletal muscle mass index (ASMI) normalized by the body size [eg, ASM/height^2] serves as an indicator of muscle mass. On the other hand, handgrip strength is a widely used and cost-effective method for evaluating muscle strength in clinical settings. It is accessible, easy to measure using a handheld dynamometer, and has shown a correlation with leg strength (Ibrahim et al., 2016; Leong et al., 2015; Rossi et al., 2014). This reliable indicator is strongly associated with clinical outcomes and commonly employed in muscle strength assessments. Measurements were taken for both the left and right hand, and the average value was calculated. The study employed two variables, muscle mass and strength, to classify participants into sarcopenic and non-sarcopenic groups. The Asian sarcopenia group's suggestions for variables were considered, but the criteria used for categorization were customized for the Korean population, considering their unique characteristics. This approach avoids adopting criteria solely based on various Asian races, as they may not entirely reflect the traits of Koreans. Therefore, in this study, the sarcopenia group was determined based on the lower and upper quantile of the entire dataset, providing a more accurate representation of sarcopenia prevalence in Koreans (Kim et al., 2018).

2.1.3. Risk factor of sarcopenia in Korea

Sarcopenia presents numerous negative health-related consequences, including impaired energy homeostasis and a detrimental impact on the quality of life for the elderly. The association between sarcopenic status in the elderly and various clinical outcomes, such as functional limitations, metabolic impairments, and increased cardiovascular risk, has become increasingly evident (Kim et al., 2015). It has been reported that individuals with severe sarcopenia are four times more likely to die within 2.6 years for various reasons, highlighting the importance of preventing sarcopenia before individuals reach old age (Bachettini et al., 2020).

Korea is one of the fastest aging countries, as indicated by the Korea National Statistical Office. The dependency ratio for individuals aged 65 and older is projected to reach 26.1% in 2023, with an estimated increase to approximately 70% over the next 20 years (KOSTAT, 2019). This ratio is expected to continue rising, leading to an increasing prevalence of sarcopenia in Korea due to the growing elderly population and increased life expectancy (Jang, 2018). Studies investigating the prevalence of sarcopenia in elderly Koreans aged 65 years and above have reported a prevalence of 14.9% in male, 11.4% in female, and an overall prevalence of 13.1% in the elderly Korean population (Choo & Chang, 2021). Furthermore, a Korean study following 500 elderly patients over the age of 65 for six years found that the mortality rate was 2.99 times higher for men and 3.22 times higher for women with sarcopenia compared to the control group (Lim et al., 2010). Hence, early prevention of sarcopenia is crucial.

2.2. Epigenetic and Sarcopenia

2.2.1. Epigenetic and DNA methylation

The term “epigenetics”, which literally means “on top of or above genetics”, refers to mitotically heritable patterns of gene expression without any change in the underlying DNA sequence. This phenomenon explains how more than 200 cell types with various functionalities and phenotypes consistently differentiate from the same DNA code (Bernstein et al., 2010; Esteller, 2006). The epigenome, on the other hand, encompasses the dynamic combination of molecular, chemical, and environmental factors that, along with the genome, determine the unique functional identity of each cell type (Jaenisch & Bird, 2003; Kanherkar et al., 2014). The physical organization of DNA, influenced by its packaging into chromatin, plays a crucial role in epigenetic gene regulation as it determines the accessibility of the genomic code to transcription factors. Epigenetic regulation of gene expression is driven by four major mechanisms: (1) cytosine methylation, (2) post-translational modifications of histone proteins such as methylation or acetylation, (3) chromatin remodeling, and (4) non-coding RNAs.

In recent years, an increasing number of publications have reported the fundamental role of epigenetic alterations in pathogenesis and disease susceptibility, including sarcopenia (Antoun et al., 2022; Feinberg, 2007; He et al., 2019). This has been a key motivation for large-scale projects that provide publicly available resources of epigenomic maps for multiple tissues, such as the Roadmap Epigenomics Project and Blueprint Epigenome Project (Adams et al., 2012). DNA methylation is an extensively studied and well-understood epigenetic mechanism that was initially proposed as a mechanism for long-term memory function (Jones, 2012). Epigenetics refers to heritable changes in gene expression that occur without altering the underlying DNA sequence. Recent advancements in next-generation sequencing and microarray technologies have greatly enhanced our understanding of epigenetics. Epigenetic mechanisms can be broadly categorized into three main types: DNA methylation, histone modification, and regulation by non-coding RNAs. DNA methylation involves the covalent addition of a methyl group to the C-5 position of the cytosine ring in DNA, resulting in the formation of 5-methylcytosine (5mC). DNA methylation is predominantly observed on cytosines followed by guanine residues (CpG), while it is less commonly found at non-CpG regions such as CpA, CpT, and CpC. CpG sites tend to cluster together in specific regions called CpG islands. In promoter regions, most CpG islands remain unmethylated to facilitate the binding of proteins to the DNA. Conversely, CpG islands located in gene bodies are often hypomethylated, leading to the silencing of repetitive DNA elements. The regulation of DNA methylation is carried out by a family of enzymes known as DNA methyltransferases (DNMTs), including DNMT1 and DNMT3. DNMT3 is responsible for de novo methylation during development, while DNMT1 acts to maintain methylation patterns during DNA synthesis (Moore et al., 2013).

2.2.2. Influences on DNA methylation

During the early stages of development, such as embryonic and fetal development, epigenetics plays a crucial role in regulating the differentiation of cells and tissues by organizing the genome into active and inactive regions for transcription. As development progresses, especially during specific critical periods of differentiation, epigenetics orchestrates the precise activation or inactivation of sets of genes that determine the mature phenotype of cells and tissues at specific developmental stages (e.g., puberty, pregnancy, aging). Abnormalities in the sequence or composition of genes regulated by epigenetic mechanisms can lead to improper gene expression, resulting in disrupted differentiation processes and the development of diseases. Epigenetics enables cells, tissues, and individuals to sense and respond to environmental cues, influencing how the genome is interpreted (Crews & Gore, 2011; Feil & Fraga, 2012; Ho, 2010; Zhang & Ho, 2011). Various external environmental factors have the potential to alter epigenetic programs in multiple cells and tissues, consequently increasing or decreasing the risk of disease development (Crews & Gore, 2011; Feil & Fraga, 2012; Ho, 2010; Zhang & Ho, 2011). In some cases, the effects of epigenetic disruptions may only become apparent at later stages of life (Tang et al., 2012; Tang et al., 2008), providing an explanation for the Barker hypothesis, which suggests that adult diseases can have their origins in early developmental stages (Godfrey & Barker, 2001; Hales & Barker, 2001).

Inheritable information carried in the primary sequence of DNA plays a major role in determining variations in susceptibility and severity of diseases. Genome-wide association studies have noticed how germline genetic variations influence disease predisposition and outcomes (Hardy & Tollefsbol, 2011; Hartman et al., 2010; Sivakumaran et al., 2011). In addition, somatic changes in DNA sequence can drastically disrupt gene expression programs, leading to the genetic progression of diseases (Hartman et al., 2010). However, in recent years, research has firmly established that genome-wide association study findings, which involve common genetic variants, alone tend not to identify causal loci of complex diseases or predict individual disease risk (Gibson, 2012).

2.3. Computational Background

2.3.1. Machine learning

In this study, performed classification of sarcopenia were performed a machine learning model, which is widely used and highly regarded machine learning model among various artificial intelligence techniques. According to Arthur Samuel (1959), machine learning is defined as "the field of study that gives computers the ability to learn without being explicitly programmed." Computers have the capability to autonomously learn without human intervention. Through machine learning algorithms, computers can construct intricate mathematical models based on a training set, enabling them to make predictions for new or unseen data. This allows computers to uncover complex patterns from high-dimensional data, a task that may be challenging for humans to process and identify. With advancements in technology and computational resources, it is now feasible to train machine learning models on extensive datasets. As a result, there has been an increasing adoption of machine learning algorithms across various fields to address real-world problems.

Machine learning encompasses various categories, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning and unsupervised learning are the main divisions based on the dependence on training data and the presence or absence of labeled data (Yu et al., 2017). Supervised learning involves training the computer using data with known labels or answers. The data is divided into training and test datasets. After training the model using the labeled training dataset, the accuracy of the trained algorithm in predicting the outcomes can be evaluated using the test dataset. For instance, when creating a model to classify images of dogs and cats, researchers provide labels for dog and cat photos in the training dataset. Once the supervised learning process is complete, the model can accurately classify new images of animals as either dogs or cats. Supervised learning is intuitive and easy to understand, but it requires the effort of researchers to obtain high-quality training data and select suitable learning algorithms to achieve good results. On the other hand, unsupervised learning is used when there are no predetermined labels in the training data. In this case, the model learns the inherent patterns within the input data without researchers specifying the groups to which the data belong. For example, when creating a model to classify animals in images without any prior information about each animal, the model learns and categorizes the animals solely based on visual patterns and colors present in the images. Unsupervised learning has the advantage of requiring less researcher intervention since there is no need to label the training data. However, it is more challenging to implement compared to supervised learning, and the results may differ from expectations due to the absence of predefined labels for the data (Arthur Samuel, 1959).

2.3.2. Machine learning in life sciences

The rapid growth of biological data has led to the increased utilization of machine learning algorithms and techniques. These methods focus on uncovering valuable insights and concealed patterns from extensive and intricate datasets, resembling the search for valuable needles within a haystack. Machine learning involves the development of computer programs that enhance their performance through experience (Mitchell, 2007). Machine learning algorithms are designed to learn from available data or a training set and optimize a performance measure. The accuracy of these optimized measures is evaluated using a fitness function that defines the optimization problem in modeling tasks. In modeling problems, the algorithm's objective is to generate a model based on the training dataset and draw conclusions from it. In optimization problems, the goal is to find an almost optimal solution among all feasible solutions, which is a fundamental aspect of modeling problems (Larranaga et al., 2006).

Machine learning method have demonstrated successful applications across various genomics domains. These include gene prediction (Mathé et al., 2002), identification of regulatory elements (Aerts et al., 2004), analysis of biological sequences, detection of functional RNAs (Won et al., 2004), and prediction of epigenetic phenomena like gene expression using epigenomic data, allele-specific DNA methylation (He et al., 2015) and DNA-based age prediction (Vidaki et al., 2017). Additionally, machine learning has been utilized in systems biology to model biological networks and in evolutionary biology to construct phylogenetic trees.

2.3.3. Use of machine learning for disease prediction

Machine learning has been widely applied across healthcare to address medical problems, including disease diagnosis (Bocchi et al., 2004; Klöppel et al., 2008), predicting health outcomes such as mortality, the development of disease, or other comorbidities (Doshi-Velez et al., 2014; Hyland et al., 2020; Yu et al., 2010) , as well as predicting adherence to treatment (Son et al., 2010) and the utilization of healthcare resources (Patel et al., 2018). Ensemble machine learning approaches (e.g. random forest) have commonly shown robust performance as classification tools in disease areas, such as Alzheimer's disease (Sarica et al., 2017) and cardiovascular disease (Hyland et al., 2020). In addition, unsupervised clustering approaches have been applied to stratify patients with similarly presenting conditions (Mossotto et al., 2017) as well as to identify different disease and comorbidity trajectories (Doshi-Velez et al., 2014).

Particularly within healthcare, data is often heterogeneous, sourced from different modalities such as questionnaires, images, recordings, and a variety of omics analyses. Although methods that are able to integrate multiple data structures for single analyses are not yet well-established, machine learning approaches used to integrate heterogeneous data have been applied to a number of areas in biomedicine (Zitnik et al., 2019).

2.3.4. Use of machine learning in sarcopenia

Supervised machine learning approaches have also been applied to predict the occurrence of sarcopenia (Gu et al., 2023), as well as the changes in future sarcopenia exacerbation (Kim, 2021) using just X-ray scans. Studies have also applied machine learning methods to optimize the management of healthcare resources, for example, by predicting activity, habits, and clinical decision-making at triage in adults presenting to the emergency department with sarcopenia exacerbation (Seok & Kim, 2023).

Although the majority of machine learning applications within the sarcopenia field have focused on X-ray medical image analysis, prediction models have also targeted whole blood analysis with clinical relevance (Kang et al., 2019). However, there are few studies analyzing DNA methylation data using machine learning. Chung, Heewon et al. (2021) used a transcriptome database from 17,339 genes from 118 subjects and accurately selected 27 genes for diagnosing sarcopenia. This study compared five machine learning algorithms against a conventional triage approach. The machine learning models selected 27 genes from three different continents (Europe, Africa, and Asia). They also created a web application to access the model. However, these improvements were not able to identify genomes that are found in Korea.

Chapter 3. Methods

3.1. Study design and population

This study was conducted a portion of the KoGES (Korean Genome and Epidemiology Study) dataset obtained from the Korean Center for Disease Control and Prevention. The largest cohort of KoGES is the survey of the Health Examinees-Gem (HEXA-G) cohort, and the dataset consists of participants' medical and pharmacological history, anthropometric traits, and blood biochemistry traits (Health Examinees Study, 2015; Kim et al., 2017) is a population-based prospective cohort consisting of 173,357 urban Korean adults who underwent health examinations at 38 medical centers. Among total of 822 participants, from whom DNA methylation data were included. KoGES_HEXA chort was used for quantitative metabolic traits analysis to examine whether metabolic traits were associated with the DNA methylation changes in identified DNA methylation CpG sites. Participants were male and female, aged 40-69 years, from 14 major cities across Korea and were recruited between 2004 and 2013.

The variables used for the criteria for sarcopenia in this study were based on the 2019 Asian Working Group on Sarcopenia (AWGS) guidelines (L. K. Chen et al., 2020). The muscle mass, assessed by appendicular skeletal muscle mass index (ASMI; ASM/hieght^2) and muscle strength, evaluated by handgrip variables were selected as diagnostic criteria. The following equation (Xu et al., 2011) was used to calculate appendicular skeletal muscle mass (ASM) as Equation 2:

$$\begin{aligned}\text{Appendicular skeletal muscle (ASM)} & \\ &= 0.193 * \text{body weight} + 0.107 * \text{height} - 4.157 * \text{gender} \\ &- 0.037 * \text{age} - 2.631\end{aligned}$$

weight in kg, height in m, age in year, gender:1 for men and 2 for women

Equation 2. Appendicular skeletal muscle mass calculator formula

The appendicular skeletal muscle mass index (ASMI) was calculated using the measured appendicular skeletal muscle mass (ASM). ASMI normalized by the body size [eg, ASM/height^2] is used as an indicator of muscle mass. Besides, muscle strength was measured by averaging the value obtained from the right and left hand. Cut-off values were applied as thresholds (Chen et al., 2014). All participants voluntarily signed an informed consent form before the study, and the study protocol was approved by the Institutional Review Boards (IRB) of the institutions that participated in KoGES. The KoGES_HEXA cohort performed in accordance with the Declaration of Helsinki and approved by the IRB of Theragen Etex (Approval Numbers: 700062-20190819-GP-006-02). And also, this article received approval from Seoul National University IRB (IRB No. E2304/003-003).

3.2. Analysis of DNA methylation arrays

Genotype data were provided by the Center for Genome Science, Korea National Institute of Health. The bisulfite-converted genomic DNA from the sample hybridized to the Illumina MethylationEPIC BeadChIP array (Illumina, San Diego, CA, USA). The raw methylation intensity data were imported using the ChAMP package (version

2.8.3)(Morris et al., 2014) in R software (Aryee et al., 2014; Morris et al., 2014). Methylation probe filtering was conducted using *champ.filter* (Zhou et al., 2017) to exclude those with at least 5% of the probes that did not pass a 0.05 detection *p*-value threshold. Probes were then filtered based on > 0.01 detection *p*-value in more than 5% of the samples and < 3 bead count in at least 5% of the samples. Exclude non-CpG probes, multi-hit probes, probes matching SNPs, and probes located in chromosomes X and Y. Quality control steps were carried out using *champ.QC* to generate various plots and dendrograms to assess the distribution of methylation. Normalization of Type-I and Type-II probes was achieved using *champ.norm* (Fortin et al., 2016; Maksimovic et al., 2012) with the peak-based correction (PBC) method. Batch correlation method *champ.combat* was applied before analyzing different methylation to avoid misestimation of cell type (Johnson et al., 2007; Leek et al., 2012). Differentially methylated DNA probes were identified using *champ.DMP* (Smyth, 2004) by comparing sarcopenia to non-sarcopenia samples, and the FDR adjusted *p*-value adjustment was applied after conducting limma analysis.

3.3. Machine-learning predictive models

Predictor selection involved identifying probes from the differentially methylated DNA probes dataset that exhibited an absolute value of 10 based logarithm of fold change (abs logFC) greater than 0.15 between sarcopenia and non-sarcopenia samples, with an FDR adjusted *p*-value less than 0.05. Both hypermethylated and hypomethylated probes that were deemed relevant were included as predictors, while the patient's binary sarcopenia and non-sarcopenia status of the patients was considered as the response variable.

3.3.1. Feature selection

Feature selection methods aim to identify a subset of the most predictive features, reducing model dimensionality and noise to improve computational demand and potentially enhance prediction accuracy. In this study, the filter and wrapper method were used and compared to identify a subset of predictors with high classification accuracy from the candidate predictors.

3.3.1.1. Pearson correlation values

The Pearson correlation is a filter method used to select a subset of relevant features. Only the features that pass the filter are included in the model that is built afterward. Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. Pearson correlation is suitable only for metric variables. The correlation coefficient has values between -1 to 1. When the value closer to 0 implies weaker correlation (exact 0 implying no correlation). And when the value closer to 1 implies stronger positive correlation and -1 implies stronger negative correlation. In this study, probes with Pearson correlation values greater than 0.7 with others were removed from the data.

3.3.1.2. Recursive feature elimination with cross-validation

The recursive feature elimination with cross-validation (RFECV) is one of the wrapper methods which needs one machine learning algorithm and uses its performances as evaluation criteria. RFECV is an iterative process that aims to identify a subset of important features. It involves evaluating the relative importance of features and removing those considered least important. Each feature is assigned a weighting and ranked based on a specified importance criterion. The feature with the lowest ranking is removed from the pool of candidate features. This process is repeated, where the remaining features are used to retrain the random forest algorithm and update the feature importance ranking. This recursive elimination continues until only a single feature remains. At each elimination step, one or multiple features can be removed (Granitto et al., 2006; Guyon et al., 2002). In this study, the RFECV method was employed following the application of the Pearson correlation method to identify the most significant probes. These selected probes were then utilized to construct the final model.

3.3.2. Classification algorithm

A classification algorithm is necessary to perform the feature selection method. In this study, voting ensemble machine learning algorithms were selected. A voting ensemble method is an ensemble machine learning model that combines the predictions form multiple other models. It is a technique that may be used to improve model performance, ideally achieving better performance than any single model used in the ensemble. A voting ensemble works by combining the predictions from multiple models. It can be used for classification or regression. In the case of regression, this involves calculating the average of the predictions from the models. In the case of classification, the predictions for each label are summed and the label with the majority vote is predicted. In this study, selected classifier performs based on the majority voting, random forest (RF) (Breiman, 2001), K-Nearest Neighbors (KNN) (Cover & Hart, 1967), Naïve Bayes (NB) (Webb et al., 2010), decision tree (DT) (Safavian & Landgrebe, 1991). The performances of the classifiers generated based on these algorithms were compared in feature selection. These algorithms are widely used to tackle various biological and medical problems. In this section and were programmed in Python via Scikit-learn module with tuning parameter.

3.3.2.1. Random forest

Random forest algorithms are utilizing a majority voting mechanism used for classification and regression tasks. They combine multiple decision trees to make predictions. Each tree is trained on a different subset of the data and a random subset of the features. During training, the trees independently make predictions, and the final prediction is determined by aggregating the individual predictions through voting (for classification) or averaging (for regression). The randomness in the selection of data and features helps to reduce overfitting and increase the model's generalization ability.

3.3.2.2. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm utilizes a majority voting mechanism and leverages data from a training dataset to make predictions for new records. When making predictions for a new record, the KNN algorithm identifies the k-closest records from the training dataset. Based on the target attribute values of these closest records, a prediction is made for the new record. The basic nearest neighbor algorithm selects the closest training instance to the arbitrary instance and returns its class label or target function value as the predicted value. The KNN algorithm extends this process by considering a specified number (k) of closest training instances instead of just one. The output of the KNN algorithm depends on whether it is used for classification or regression. In classification, the predicted class label is determined by majority voting among the selected k instances. In regression, the predicted value is the average of the target function values of the nearest neighbors. The choice of k allows for balancing between overfitting prevention and resolution, where higher values of k can help prevent overfitting but may result in less differentiation among similar instances.

3.3.2.3. Naïve Bayes

Naïve bayes (NB) is a probabilistic classification algorithm based on Bayes' theorem. It assumes that all features are independent of each other given the class variable. The algorithm calculates the probability of a class label based on observed features using prior probabilities and likelihood. It selects the class label with the highest posterior probability as the prediction. NB algorithm is simple, efficient, and widely used in text classification and spam filtering. It requires a small amount of training data and is computationally efficient. However, it may not perform well if the independence assumption is strongly violated or if there is a significant class imbalance. Overall, NB algorithm is a powerful algorithm for fast and reliable classification tasks, leveraging probabilistic principles and feature independence assumption.

3.3.2.4. Decision tree

Decision tree algorithms are versatile and widely used in machine learning for classification and regression tasks. It uses a majority voting mechanism to make predictions. The algorithm builds a tree-like model where each internal node represents a feature, and each leaf node represents a class label or a predicted value. The algorithm splits the data based on feature values to create homogeneous subsets. At each internal node, the algorithm selects the feature that provides the best split, often using measures like information gain or Gini impurity. The majority voting mechanism comes into play when multiple samples reach a leaf node with different class labels. In classification, the class label that occurs most frequently among the samples is chosen as the prediction. In regression, the average value of the target variable among the samples is used. Decision trees are intuitive, interpretable, and capable of capturing complex relationships. However, it can be prone to overfitting and are sensitive to small changes in the data. Therefore, decision tree is commonly used to mitigate these issues by combining multiple decision trees and aggregating their predictions through majority voting.

3.3.3. Hyperparameter tuning

Each of the described machine learning algorithms has a set of hyperparameters that can be tuned to optimize classifications based on a given training dataset. However, not all hyperparameters are equally important for tuning purposes (Bergstra & Bengio, 2012). Grid search is a hyperparameter search strategy that exhaustively explores all combinations of hyperparameters to find the optimal set based on a specified performance measure within a cross-validation framework. While grid search is widely used, it can be computationally expensive depending on the algorithm and the size of the hyperparameter space. Moreover, considering a large number of hyperparameters may compromise the performance of grid search (Bergstra & Bengio, 2012). In contrast, random search involves randomly selecting a specified number of hyperparameter combinations to evaluate. It is a faster and less computationally demanding method compared to grid search because it does not exhaustively search all possible combinations. However, there is a possibility of missing the optimal hyperparameter set. Nevertheless, studies have shown that random search is effective for evaluating a large hyperparameter search space, focusing on important tuning parameters, and often able to identify the optimal hyperparameter set (Bergstra & Bengio, 2012). When dealing with large parameter spaces, a dual search strategy can be employed to combine the advantages of random and grid search approaches. In this dual search strategy, a random search is initially conducted to quickly evaluate and narrow down the large hyperparameter search space. Then, an exhaustive grid search is applied to the condensed search space to definitively identify the best hyperparameter set (Bergstra & Bengio, 2012).

In this research, the hyperparameters of the four machine learning algorithms were adjusted to optimize their performance. However, it was worth mentioning that hyperparameter tuning was conducted exclusively during the model training phase, as the default parameters are already known to deliver satisfactory predictive accuracy across a wide range of problem scenarios.

3.4. Statistical analysis

Three types of analysis tools were used in this study. First, to analysis participants characteristics SPSS statistical program (Ver 26.0.02 Chicago IL, USA) was used to analyze the data as follows.

1) All demographic items were analyzed using descriptive statistics to obtain the mean (M) and standard deviation (SD).
2) The significance level of all statistics is set at p<.05.

Second, differentially methylated DNA probes data obtained through statistical analysis were analyzed using R studio (Ver 2022.12.0 + 353) from DNA raw data. Lastly to analyze the learning process of each machine learning model using Scikit-learn: Machine Learning in Python (Pedregosa et al., 2011) (Ver 3.9.13) was utilized.

Chapter 4. Results

4.1. Clinical characteristics of study participants

The clinical characteristics of the study participants are presented in Table 1. A total of 110 individuals were included in the analysis to investigate the DMPs associated with sarcopenia, comprising 82 male (37 sarcopenia cases and 45 non-sarcopenia controls) and 28 female (12 sarcopenia cases and 16 non-sarcopenia controls) were included to investigate the DMPs associated with sarcopenia.

Among the male participants, the non-sarcopenia group exhibited higher values for weight, body mass index, handgrip strength, and appendicular skeletal muscle mass index (ASMI) compared to the sarcopenia group. However, significant differences were observed for each variable in the female participants.

4.2. Identification differentially methylated probes of sarcopenia

The study procedures for the identification of sarcopenia diagnostic biomarkers are presented in Figure 1. Preprocessing steps were performed, resulting in 521,616 final probes for further analysis after removing probes with missing values, SNP overlap, or locating on sex chromosomes, from the initial 740,236 probes in the primary data. In male group, a total of 740,236 CpG sites were investigated for their association with sarcopenia. And in female group, 740,082 CpG sites were examined. In male group found 166 DMPs ($|\mathrm{logFC}| > 0.15$ and

FDR adjusted p-value $<$ 0.05), which consisted of 99 and 67 hyper and hypomethylated positions, respectively (Figure 2). In contrast, the female group did not meet the threshold of FDR adjusted p-value $<$ 0.05 for the DMPs, preventing further analysis in this cohort. Therefore, the data for the female group could not be included due to the lack of significant findings in relation to the DMPs.

In male group, the distribution of these male DMPs in relation to CpG islands (CGIs) and genomic regions such as North Shelf, North Shore, CGI, South Shores, South Shelf, and Open Sea were explored. Hypermethylated sites were predominantly found in CGI, and their number generally decreased with increasing distance from CpG islands (Figure 3(A)). Additionally, the distribution of DMPs in relation to transcription start sites (TSSs), including TSS1500, TSS200, 5' untranslated region (5'UTR), first exon (1st Exon), and gene body, was examined. Hypermethylated probes were primarily observed in TSS1500, while hypomethylated sites were predominantly located in the 5'UTR and first exon (Figure 3(B)).

4.3. Identification of diagnostic biomarkers for sarcopenia using machine learning

In order to identify diagnostic biomarkers for sarcopenia in male group, feature selection was performed on a dataset consisting of 166 DMPs, which included all available candidate predictors. The potential presence of multicollinearity was evaluated by assessing the correlation between features within the selected subsets, using

Pearson's correlation coefficient. Among the candidate probes, those with a Pearson correlation coefficient greater than or equal to 0.7 were excluded. This correlation-based filtration process resulted in the removal of 135 probes from the initial set of candidates. Feature selection using the RFECV method was then conducted. This method identified optimal subsets of 10 features, as shown in Table 4, achieving an average cross-validation balanced accuracy score of 95% for the sarcopenia different methylation models (Figure 4).

4.4. Complete machine learning model development

Complete data for the 10 selected probes for the sarcopenia model was available. Following the stratified train-test split, 70% of the dataset was allocated to the initial training set, while the remaining 30% was assigned to the test set. All four machine learning classifiers trained on the complete training dataset exhibited moderate discriminative performance in the training set, with AUC values ranging between 0.86 and 1.0. Lastly, voting classifier trains different models using the chosen algorithms, returning the majority's vote as the classification result. Among these classifiers, the random forest model and Naïve Bayes model showed the highest sensitivity, achieving 100%. Among them, the AUC was found to be 98% by majority voting, which suggests that the prediction model with 10 feature selections performs well (balanced accuracy, sensitivity, and F1-score; illustrated Table 3).

4.5 Exploration of model optimization techniques

Based on the performance of the sarcopenia dataset developed using the complete training datasets, the best-performing model was selected to evaluate whether addressing the extent of missing data and class imbalance present in the training datasets could enhance the predictive performance of the developed models. The selection of the best-performing model was primarily based on the AUC criterion (refer to Figure 6).

4.6. Correlation with age using final biomarker of sarcopenia

Utilizing the 10 most important probes from the male group (Table 4), further analysis was performed on 10 probes associated with age, taking into consideration the relationship between sarcopenia and aging (refer to Figure 7). Among 10 probes RYR2 (cg00299070) showed the highest negative correlation with age (r = −0.64). cg00008452 shows the lowest negative correlation with age (r = -0.45). cg08906030 shows the highest positive correlation with age (r = 0.59). cg0016066 shows the lowest positive correlation with age (r = 0.32).

Table 1. Characteristics of study participants

(A) Male participants characteristics

	Sarcopenia (n=37)	Non-sarcopenia (n=45)	*p*-value
Age (yrs)	52.9 ± 5.8	45.8 ± 4.9	.359
Height (cm)	166.7 ± 5.2	172.5 ± 6.1	.527
Weight (kg)	57.4 ± 4.0	94.5 ± 8.3	.001**
BMI (m^2/kg)	20.6 ± 1.0	31.7 ± 1.6	.005**
Handgrip strength (kg)	28.2 ± 3.7	48.3 ± 5.0	.010*
ASMI (ASM/m^2)	7.3 ± 0.2	9.5 ± 0.3	.004**

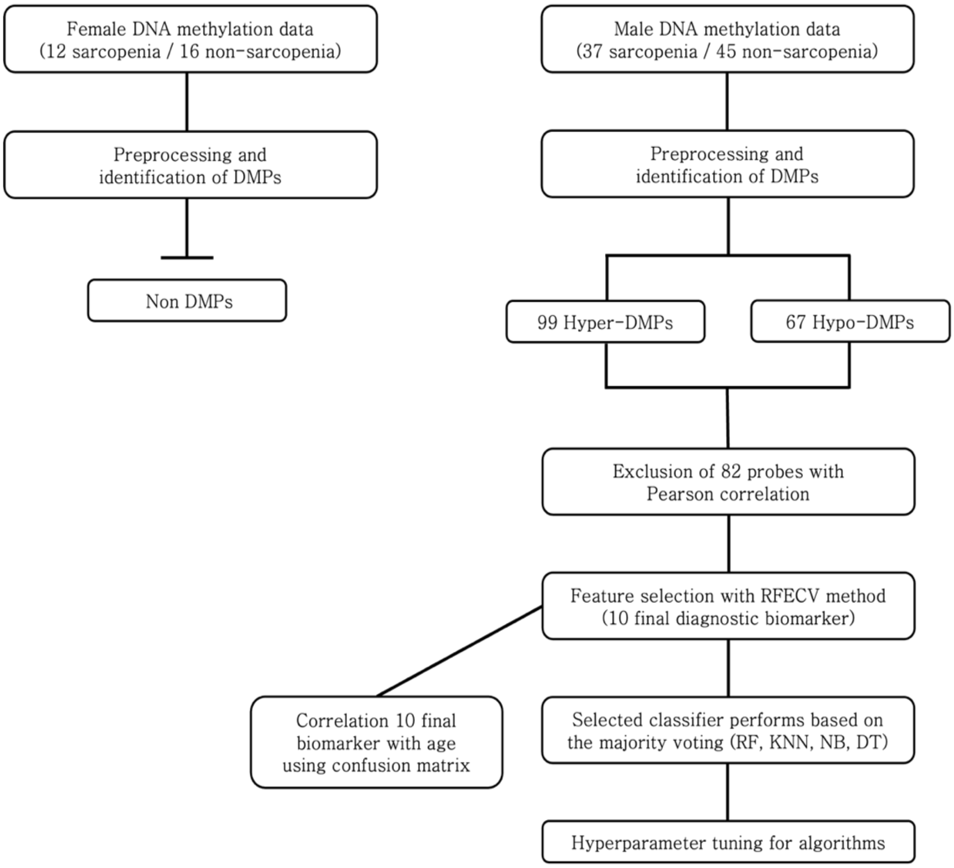
(B) Female participants characteristics

	Sarcopenia (n=37)	Non-sarcopenia (n=45)	*p*-value
Age (yrs)	55.0 ± 2.9	52.1 ± 2.2	.004**
Height (cm)	151.8 ± 6.5	159.0 ± 4.4	.001**
Weight (kg)	48.9 ± 3.6	65.9 ± 3.4	<.001**
BMI (m^2/kg)	21.2 ± 1.5	26.1 ± 0.8	<.001**
Handgrip strength (kg)	16.1 ± 1.9	27.6 ± 2.2	<.001*
ASMI (ASM/m^2)	5.5 ± 0.2	6.7 ± 0.1	<.001**

Data are presented in mean ± standard deviation (SD) or median (interquartile range [IQR]). Statistical difference analysis was performed with t test for continuous variables and chi-square test for categorical variables. *p<.05, **p<.01

Abbreviations: *BMI,* body mass index; *ASMI,* appendicular skeletal muscle index; *ASM,* appendicular skeletal muscle

Figure 1. Flow chart



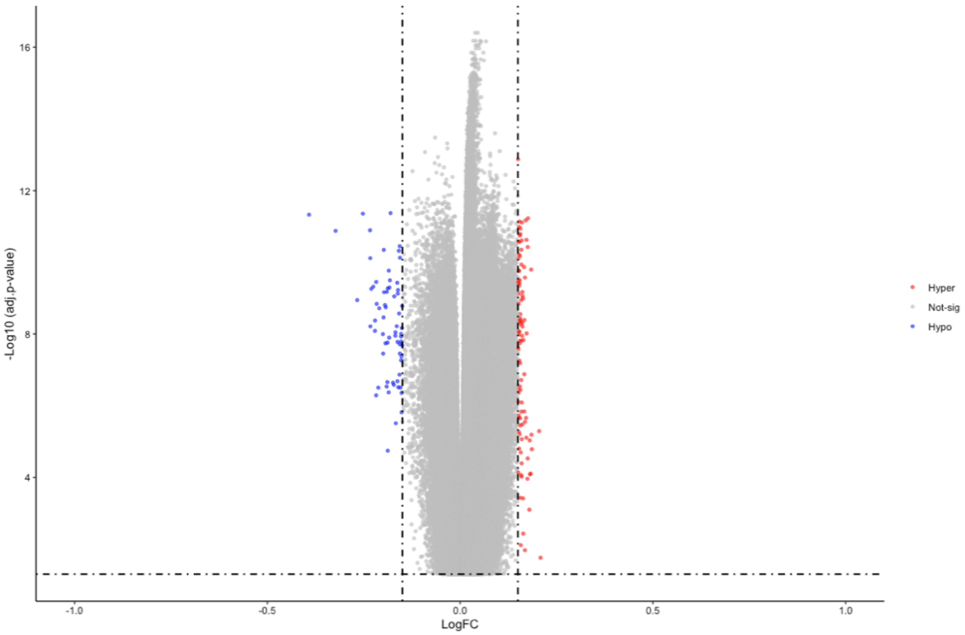
Female DNA methylation data (12 sarcopenia / 16 non-sarcopenia)
Preprocessing and identification of DMPs
Non DMPs
Male DNA methylation data (37 sarcopenia / 45 non-sarcopenia)
Preprocessing and identification of DMPs
99 Hyper-DMPs
67 Hypo-DMPs
Exclusion of 82 probes with Pearson correlation
Feature selection with RFECV method (10 final diagnostic biomarker)
Correlation 10 final biomarker with age using confusion matrix
Selected classifier performs based on the majority voting (RF, KNN, NB, DT)
Hyperparameter tuning for algorithms

The workflow for identifications of sarcopenia diagnostic biomarkers. Male and female group are analysis respectively. However, female group could not meet criteria of differentially methylated probes so this study analysis male group.

Hypermethylated and hypomethylated probes locating in the promoter region of genes were used for biomarker selection. After removing correlated features, 10 final methylation probes were identified using feature selection. Different algorithms were compared based on their accuracy and combining similar different machine learning classifiers to classification via majority voting. Each algorithm's hyperparameters were tuned and its performance was evaluated. In order to examine the relationship between age and the expression levels of 10 probes, a comprehensive analysis using a confusion matrix was conducted.

Abbreviations: *DMPs*, differentially methylated probes; *RFECV*, recursive feature elimination with cross-validation; *RF*, random forest; *KNN*, K-Nearest Neighbors *NB*, Naïve Bayes; *DT*, Decision Tree

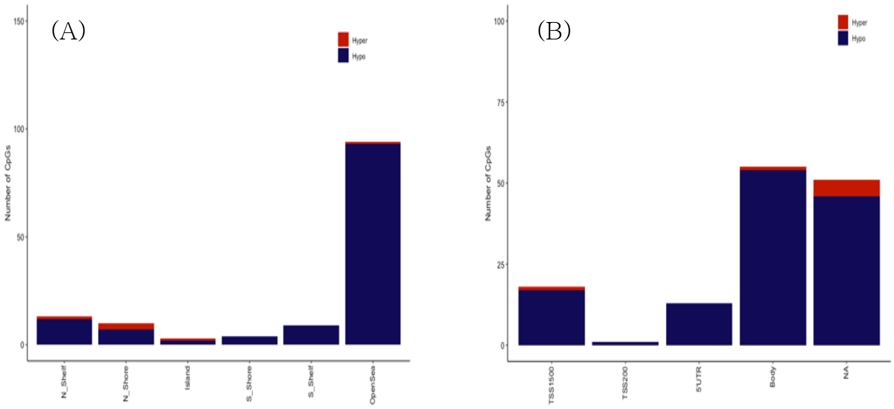
Figure 2. Differentially methylated probes between in male group



Male group DMPs were found with the criteria of |logFC | > 0.15 and adjusted FDR *p*- value < 0.05. Red and blue dots are hypermethylated and hypomethylated DMPs, respectively. And the gray ones were not significant according to the above-mentioned criteria. Among 166 DMPs 99 (almost 60%) were hypomethylated (shown in blue) and 67 (almost 40%) were hypermethylated (shown in red) in the severe group, and the grey ones were not significant according to the defined criteria.

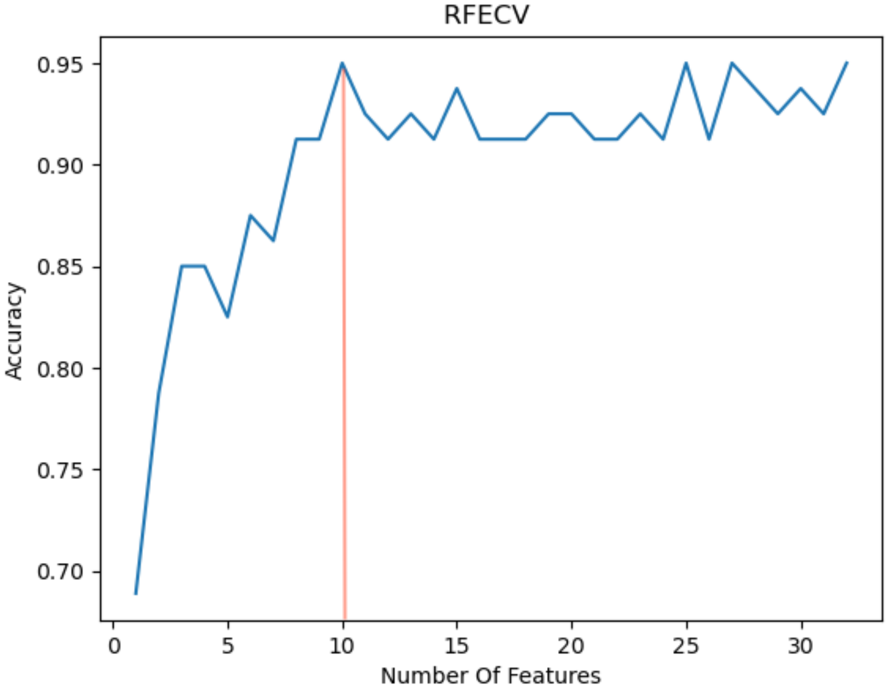
Abbreviations: *adj-p-value*, Benjamini/Hochberg adjusted *p*-value; *Log FC*, Log Fold change; *Hyper,* Hyper-methylation; *Hypo,* Hypo methylation

Figure 3. Identification of sarcopenia differentially methylated probes location in male groups



Distribution of DMPs in differentially methylated probes defined by the distance from CGI. Red and dark blue parts stand for hypermethylated and hypomethylated DMPs, respectively (A). Distribution of DMPs in different genomic regions defined by the distance from TSS. Red and dark blue parts represent for hypermethylated and hypomethylated DMPs, respectively (B). Abbreviations: *1stExon*, the first exon; *5´UTR, 5´* untranslated region; *adj-p-value*, Benjamini/Hochberg adjusted *p*-value; *CGI*, CpG island; *DMPs*, differentially-methylated probes; *Island*, CpG island; *N_Shelf*, North Shelf (2-4 kb upstream of CGI); *N_Shore*, North Shore (0-2 kb upstream of CGI); *Open Sea*, further than 4 kb from CGI; *S_Shelf*, South Shelf (2-4 kb downstream of CGI); *S_Shore*, South Shore (0-2 kb downstream of CGI); *TSS*, transcription start site; *TSS1500*, 200-1500 nucleotides upstream of TSS); *TSS200*, 0-200 nucleotides upstream of TSS.

Figure 4. Identification of sarcopenia probe in male group using feature selection



Feature selection with RFECV. Red line represents the optimal subset of features for inclusion in each model was identified as the subset which offered the best-balanced accuracy score (95%). Checking different number of features on X axis and validation score on Y axis.
Abbreviations: *RFECV*, Recursive Feature Elimination with Cross Validation

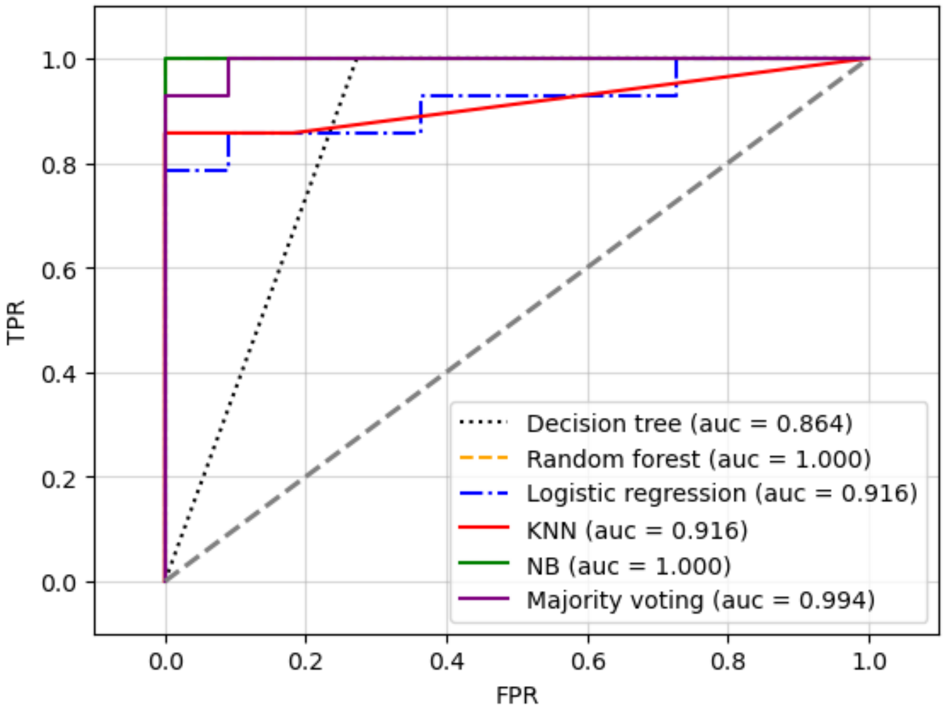
Table 2. Comparison of prediction performances in test dataset

Model	Precision	Recall	Accuracy	F1	AUROC
Decision tree	0.8235	1.0000	0.8800	0.9032	0.8500 (+/- 0.200)
Random forest	1.0000	0.8571	0.9200	0.9231	1.0000 (+/- 0.000)
KNN	0.9231	0.8571	0.8800	0.8889	1.0000 (+/- 0.000)
NB	1.0000	0.9286	0.9600	0.9630	1.0000 (+/- 0.000)
Majority Voting	1.0000	0.9286	**0.9600**	0.9630	1.0000 (+/- 0.000)

Summarizes the comparison of the cross-validation accuracy. The results show that majority voting provided the highest accuracy metrics.

Abbreviations: *KNN*, K-nearest neighbors; *NB*, Naïve Bayes; *F1*, F1 score; *AUROC*, area under the receiver operating characteristics

Figure 5. ROC curves comparing the performance of classifiers



ROC curves comparing the performance of all classifiers. From the ROC curve, can see that the majority voting ensemble classifier performs well on the test set.
Abbreviations: *TPR*, True Positive Rate; FPR, False Positive Rate; *KNN*, K-Nearest neighbor; *NB*, Naïve bayes

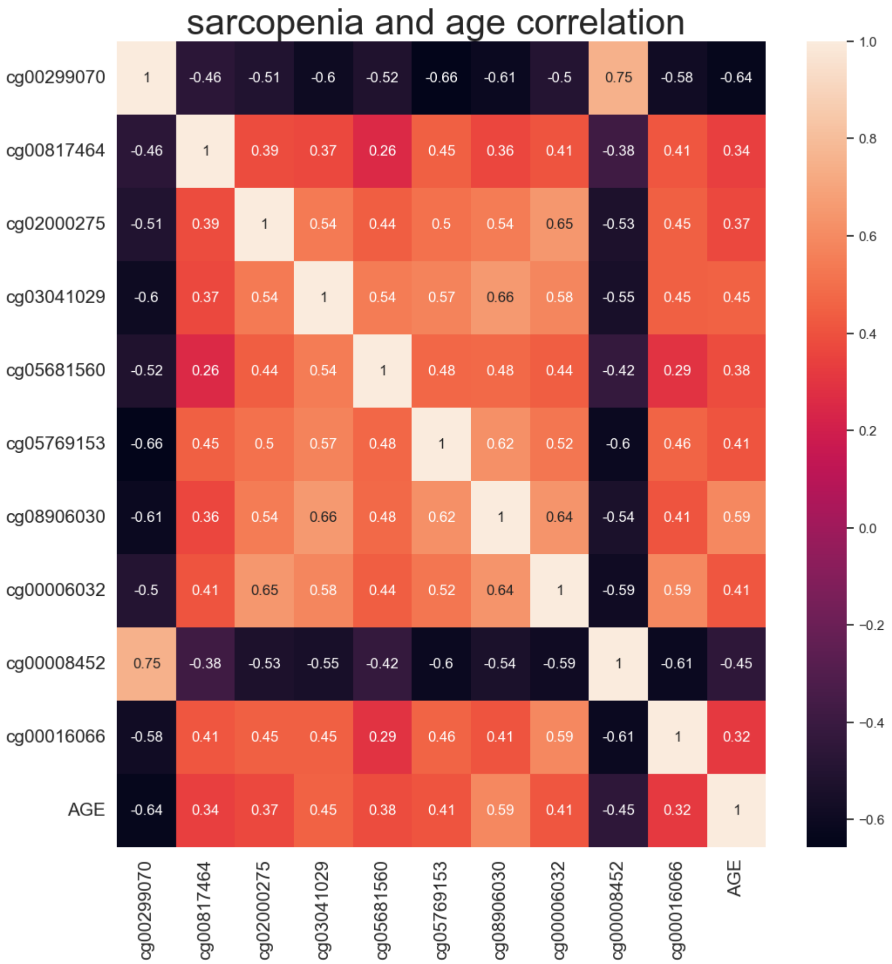
Table 3. 10 selected probes for the diagnostic model

Probe	Chromosome	Gene	Genomic position
cg00299070	chr 01	RYR2	Body-shore
cg00817464	chr 10	XPNPEP1	Body-opensea
cg02000275	chr 14		IGR-opensea
cg03041029	chr 02		IGR-opensea
cg05681560	chr 06		IGR-opensea
cg05769153	chr 19	ZNF415	TSS1500-shore
cg08906030	chr 16		IGR-shore
cg00008452	chr12	APAF1	Body-opensea
cg00006032	chr 14	GPHN	1st Exon-island
cg00016066	chr 15	TCF12	TSS200-island

Shows the features of these 10 final probes, the positions of the 10 probes, and the names of the genes. Gene names not mentioned in this list represent undiscovered genes. The probes ⊿meth > 5%.

Abbreviations: *CHR*; chromosome, *CGI*; CpG island, Based on Genomic position; opensea, shore, island, shelf, *IGR*; Intergenic Region, *OpenSea*; further than 4 kb CpG island, *Shore*: further than 2 kb CpG island, *1st Exon*; Initial exon of a gene, *TSS*; Transcription Start Site of a gene; *⊿meth*; delta methylation beta value

Figure 6. Heatmap depicting the correlation between sarcopenia and age



The generated heat map illustrates the correlation between DNA methylation levels of 10 probes and age, providing a visual representation of the relationship between DNA methylation patterns and chronological age. cg00299070 shows that highest negative correlation with age.

Chapter 5. Discussion

5.1. Summary of findings

The decline in muscle mass and function among Korean older adults poses a significant health concern, underscoring the importance of early prediction and diagnosis of sarcopenia to facilitate timely intervention. However, the definition of sarcopenia has only been recently established, and there is no universally accepted standard for its diagnosis. Hence, it becomes crucial to establish an early diagnostic screening model that can identify characteristic biomarkers associated with sarcopenia. Instead of focusing on phenotypic diagnosis, this study opted to develop a gene level diagnostic model for sarcopenia, leveraging the advantages of machine learning methods in gene selection and classification. Recent advancements in machine learning techniques, coupled with the availability of gene expression data in public databases, have opened up new diagnostic and predictive possibilities for sarcopenia.

Considering previous evidence demonstrating increased DNA methylation in muscles due to aging or diseases, it was hypothesized that sarcopenia might exhibit alterations in DNA methylation. While some studies have reported an overall rise in DNA hypermethylation with age in muscle tissue (Day et al., 2013; Turner et al., 2020; Zykovich et al., 2014), indicating specific genomic regions or genes manifest higher methylation levels in older individuals. However, this study provides a unique insight, revealing a slight decline in DNA hypomethylation with age (Blocquiaux et al., 2022), suggesting lower methylation levels in older individuals. Based on these findings, this study selected both hypomethylation and hypermethylation probes to investigate the different factors contributing to sarcopenia.

In this research, gathered microarray expression profiling datasets from the KoGES database, identifying 166 DMPs distinguishing between sarcopenia and non-sarcopenia blood samples. Among these DMPs, employed the Pearson correlation method to assess correlation coefficients and filter out irrelevant variables, followed by using the RFECV method to select significant variables. Subsequently, identified 10 signature probes that exhibited differences between the sarcopenia and non-sarcopenia groups. The voting algorithm was then utilized to assess the diagnostic accuracy of these 10 essential probes. By comparing the prediction accuracy of the constructed diagnostic model on the training and test datasets using the AUC of the ROC curve, observed that the model exhibited strong diagnostic capability. To the best of knowledge, this is the first study to construct a diagnostic model for Korean sarcopenia.

5.2. Correlation with different methylation CpG sites with sarcopenia

A comparison between the muscle transcriptome of older individuals with sarcopenia and healthy age-matched controls revealed that mitochondrial dysfunction was the predominant transcriptional signal associated with sarcopenia (Migliavacca et al., 2019). Moreover, this study identified oxidative phosphorylation as one of the pathways enriched among the differentially methylated CpG sites (dmCpGs) associated with sarcopenia, indicating that altered DNA methylation might play a role in mediating or consolidating the observed transcriptional changes in sarcopenic muscle.

However, out of the 166 samples with methylation data, only a small percentage (9.7% of dmCpGs and 2.2% of CpGs within differentially methylated regions) exhibited a significant correlation between DNA methylations. This suggests that a considerable number of CpGs may need to undergo differential methylation before any changes in gene expression occur, highlighting the complex relationship between DNA methylation in sarcopenia.

5.3. Selection of genomic markers

The development of sarcopenia is intricate and involves the regulation of multiple cellular signaling pathways. Generally, essential genes, signal transduction pathways, and protein-protein interaction (PPI) networks work together to maintain a balance between muscle protein synthesis and degradation, influenced by cues related to muscle hypertrophy and atrophy. In this research, pinpointed DMPs linked to sarcopenia through gene expression differential analysis. Subsequently, identified ten critical DMPs using a random forest classifier and crafted a unique diagnostic model for sarcopenia in Koreans using a voting ensemble approach, marking the first time such a model has been created for this population.

The 10 probes extracted from the RFECV method showing discriminative power between sarcopenia and non-sarcopenia samples included cg00299070 (RYR2), cg00817464 (XPNPEP1), cg02000275, cg03041029, cg05681560, cg05769153 (ZNF415), cg08906030, cg00008452 (APAF1), cg00006032 (GPHN), and cg00016066 (TCF12). According 10 probes with the highest accuracy performance is RYR2. Ryanodine receptor 2 (RYR2) is a ryanodine receptor, a major cellular mediator of calcium-induced calcium release (CICR) in animal cells and forms a class of intracellular calcium channels in various types of excitable animal tissues (Marx & Marks, 2013). Overexpression of RYR2 causes neurodegenerative diseases, heart failure, cardiac arrhythmias, and diabetes (Marks, 2023). The over expression of RYR2 in skeletal muscle of the elderly suggests that it may contribute to cardiac or diabetic disease and, consequently, to the development of sarcopenia. Previous studies have also shown that RYR2 can cause sarcopenia through its association with RYR1 (Bauerová-Hlinková et al., 2020), which shares the same receptor as RYR2. Interestingly, unlike RYR2, dysfunction of RYR1 contributes to a variety of muscle dysfunctions, including muscle weakness, age-related loss of muscle function, and cancer-related muscle weakness. The two receptors interact to cause a defect in the release of Ca2+ ions from the endoplasmic reticulum into the myocyte cytoplasm. This may suggest that overexpression of RYR2 may cause mutations in RYR1 that affect muscle cells.

In a study conducted by Wang, S., et al. (2022), it was demonstrated that the deletion of TCF12 leads to a reduction in myofiber size during muscle development, resulting in a decrease in muscle mass. The inducible deletion of TCF12 in muscle stem cells was found to cause a delay in muscle regeneration. Additionally, the inducible deletion of TCF12 in adult mesenchymal stem cells was observed to have a similar effect on muscle regeneration. Further examination of muscle stem cells revealed that the deletion of TCF12 resulted in cell-autonomous defects during muscle formation and that TCF12 played a crucial role in regulating muscle gene expression. Mechanistically, it was discovered that TCF12 and MYOD worked together to stabilize chromatin conformation and maintain the expression of genes related to muscle cell fate commitment and chromatin structure factors. In conclusion, TCF12 was identified as an important regulator of MuSC (muscle stem cell) chromatin remodeling, playing a role in controlling muscle cell fate and contributing to skeletal muscle development and regeneration. This study also found that these genes were hypomethylated, which may suggest lower TCF12 expression in sarcopenia patients.

ZNF415 is belongs to a category of protein-coding genes known as transcription factors. These transcription factors act to regulate gene expression by binding to specific DNA sequences. The gene was identified in a previous research paper aimed at identifying potential key biomarkers for sarcopenia in Americans and developing an early diagnostic model (Lin et al., 2022). The exact function and role of ZNF415 has not been studied as extensively as other genes, but its identification in this and previous studies suggests that it may be involved in a variety of cellular processes that may affect sarcopenia.

5.4. Prediction generalizability, robustness and resolution

Despite the availability of sarcopenia criteria specifically designed for Asian populations, many healthcare providers remain unfamiliar with the assessment and diagnosis of sarcopenia. Also lack of a clear definition of sarcopenia in the Korean population poses a limitation in epidemiologic studies (Baek et al., 2023). In this study adopted a definition of sarcopenia based on strength and muscle mass, with the aim of encompassing individuals at risk of developing sarcopenia. Although the model developed in this study shows potential in the diagnosis of sarcopenia in Korean male, but performance has not yet been validated in real-life sarcopenia patients. The primary objective of this study was to leverage machine learning techniques to construct a model capable of accurately classifying genomes associated with sarcopenia in the Korean population. Subsequent research endeavors could employ this model to identify individuals at risk of sarcopenia, enabling early implementation of primary or secondary prevention and management strategies.

To promote the clinical use of complex machine learning methods, studies must address the major hurdle of model interpretability. Attempts have been made by studies such as Bose et al. to address the issue of model interpretability using feature importance measures that generate an importance ranking for the predictors included in the model. However, feature importance is limited in that it is unable to offer insight into the direction of the predictor's effect or provide information on how predictors interact to deduce individual predictions. By using two types of feature selection method, such information was extractable from the sarcopenia data machine learning models and enabled both global and local explanations of model predictions to be uncovered.

5.5. Strength and limitation

In the current study, utilizing machine learning approached on KoGES data, identified a set of ten methylation probes, being differentially methylated probes in sarcopenia compared with non-sarcopenia whole blood samples, which can be used in future for the diagnosis sarcopenia in Korean population. Also, the results showed that these models outperformed equivalent models developed using voting methods, leading to more robust predictions. Further experimental evaluations and clinical validations are needed for the approval of these candidate biomarkers.

However, this study has several notable limitations. Firstly, the dataset was limited to Korean only, and the results may differ in other demographic populations. Additionally, the data used in the study was slightly outdated, covering the period from 2004 to 2013. Although the outcomes may not vary significantly for other years, it is important to acknowledge the potential influence of using older data. Since the KoGES data has been collecting follow-up data since 2013, there is a possibility of utilizing this sarcopenia model for studying methylation risk scores in the future.

Secondly, the prediction model for sarcopenia had a limited number of samples. Machine learning models generally benefit from larger datasets, and incorporating more data would likely improve the performance of the developed models. To maintain an appropriate sample size for machine learning, feature selection was performed prior to conducting a train-test split. However, this approach may have resulted in information leakage, and it should be noted that the model lacked a validation set. A valid alternative would have been to utilize nested cross-validation, which allows for the use of all available data for training and testing, while obtaining confidence intervals to assess the generalizability of the predictions. Furthermore, since there is no universally agreed-upon threshold for sarcopenia prediction models, the performance measures were evaluated using the classification threshold that maximized the Youden's index. While this aligns with current study methods, it is important to acknowledge that the choice of threshold cutoff for classification can impact the reported performance. Therefore, until a consensus is reached within the clinical community regarding the most appropriate threshold to use, performance measures derived from the confusion matrix should be evaluated while considering the variation in thresholds employed across studies.

Thirdly, DNA methylation was measured in blood samples. Although blood samples are commonly used for DNA methylation analysis, the performance could differ when using sarcopenia-relevant tissues, such as skeletal muscle. It is well-known that DNA methylation is significantly influenced by the tissue type. While peripheral blood-based DNA has lower specificity compared to DNA from disease-relevant tissue due to its mixture of DNA from various cells

(Christensen et al., 2009; Lokk et al., 2014; Widschwendter et al., 2008), it is more accessible for collection due to better patient compliance. Some studies suggest that epigenetic changes in DNA from peripheral blood can reflect changes in tissues (Teschendorff et al., 2009; Woodson et al., 2001), and statistical methods were employed in this study to calculate cell type composition and minimize the confounding effect of cellular heterogeneity (Houseman et al., 2012).

Finally, it should be noted that this model considered only clinical, environmental, and simple biomarker predictors. Incorporating genomic predictors could potentially further improve sarcopenia predictions. However, the aim of this study was to explore whether machine learning methods could outperform existing logistic regression models, which appeared to have limitations. Therefore, at this stage of the study, sarcopenia genomic biomarkers were not included to provide a fair comparison with the existing regression-based models.

Chapter 6. Overall summary and Conclusion

In summary, this study identified 166 differentially methylation probes associated with sarcopenia, out of which 10 core probes (RYR2, XPNPEP1, cg02000275, cg03041029, cg05681560, ZNF415, cg08906030, APAF1, GPHN, TCF12) were singled out by the machine learning algorithms. Notably, three of these genes (RYR2, TCF12, and ZNF415) encoded proteins that hold potential as sarcopenia biomarkers. RYR2 was considered a possible biomarker due to its involvement in ryanodine receptors found in heart muscle endoplasmic reticulum and its interaction with RYR1 receptors associated with muscle diseases. TCF12 also encodes a protein linked to delayed muscle regeneration, and its hypomethylation suggested an important role in regulating muscle gene expression, especially during muscle formation. While ZNF415 is still not fully understood compared to other genes, previous studies have also identified it in diagnostic models of sarcopenia, hinting at its significance in gene methylation regulation through DNA sequence binding associated with sarcopenia.

This study also discovered a strong correlation between age and the 10 identified probes. Particularly, RYR2, which is known for its association with muscle disorders, showed age dependency. These findings provide valuable insights into the molecular mechanisms underpinning sarcopenia and suggest the relevance of age-related DNA methylation changes in the pathogenesis of this condition.

Overall, this research successfully identified several potential genetic biomarkers and developed a high-performance early prediction model for sarcopenia in Koreans. The study results are anticipated to provide a valuable reference for future early diagnosis and screening of sarcopenia.

Reference

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E. T., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., . . . Willcocks, S. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, *30*(3), 224-226.

Antoun, E., Garratt, E. S., Taddei, A., Burton, M. A., Barton, S. J., Titcombe, P., Westbury, L. D., Baczynska, A., Migliavacca, E., & Feige, J. N. (2022). Epigenome-wide association study of sarcopenia: findings from the Hertfordshire Sarcopenia Study (HSS). *Journal of Cachexia, Sarcopenia and Muscle*, *13*(1), 240-253.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, *30*(10), 1363-1369.

Bachettini, N. P., Bielemann, R. M., Barbosa-Silva, T. G., Menezes, A. M. B., Tomasi, E., & Gonzalez, M. C. (2020). Sarcopenia as a mortality predictor in community-dwelling older adults: a comparison of the diagnostic criteria of the European Working Group on Sarcopenia in Older People. *European journal of clinical nutrition, 74*(4), 573-580.

Baek, J. Y., Jung, H. W., Kim, K. M., Kim, M., Park, C. Y., Lee, K. P., Lee, S. Y., Jang, I. Y., Jeon, O. H., & Lim, J. Y. (2023). Korean Working Group on Sarcopenia Guideline: Expert Consensus on Sarcopenia Screening and Diagnosis by the Korean Society of Sarcopenia, the Korean Society for Bone and Mineral Research, and the Korean Geriatrics Society. *Annals of Geriatric Medicine and Research, 27*(1), 9-21.

Barrès, R., Yan, J., Egan, B., Treebak, J. T., Rasmussen, M., Fritz, T., Caidahl, K., Krook, A., O'Gorman, D. J., & Zierath, J. R. (2012). Acute exercise remodels promoter methylation in human skeletal muscle. *Cell methabolism*, *15*(3), 405-411.

Bauerová-Hlinková, V., Hajdúchová, D., & Bauer, J. A. (2020). Structure and Function of the Human Ryanodine Receptors and Their Association with Myopathies-Present State, Challenges, and Perspectives. Molecules (Basel, Switzerland), 25(18), 4040.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045-1048.

Blocquiaux, S., Ramaekers, M., Van Thienen, R., Nielens, H., Delecluse, C., De Bock, K., & Thomis, M. (2022). Recurrent training rejuvenates and enhances transcriptome and methylome responses in young and older human muscle. *JCSM Rapid Communications*, *5*(1), 10-32.

Bocchi, L., Coppini, G., Nori, J., & Valli, G. (2004). Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. *Medical engineering & physics*, *26*(4), 303-312.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, L.-K., Liu, L.-K., Woo, J., Assantachai, P., Auyeung, T.-W., Bahyah, K. S., Chou, M.-Y., Chen, L.-Y., Hsu, P.-S., & Krairit, O. (2014). Sarcopenia in Asia: consensus report of the Asian Working Group for Sarcopenia. *Journal of the American Medical Directors Association*, *15*(2), 95-101.

Chen, L.-K., Woo, J., Assantachai, P., Auyeung, T.-W., Chou, M.-Y., Iijima, K., Jang, H. C., Kang, L., Kim, M., & Kim, S. (2020). Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. *Journal of the American Medical Directors Association*, *21*(3), 300-307. e302.

Chen, L. K., Woo, J., Assantachai, P., Auyeung, T. W., Chou, M. Y., Iijima, K., Jang, H. C., Kang, L., Kim, M., Kim, S., Kojima, T., Kuzuya, M., Lee, J. S. W., Lee, S. Y., Lee, W. J., Lee, Y., Liang, C. K., Lim, J. Y., Lim, W. S., . . . Arai, H. (2020). Asian Working Group for Sarcopenia: 2019 Consensus Update on Sarcopenia Diagnosis and Treatment. *Journal of the American Medical Directors Association*, *21*(3), 300-307.e302.

Choo, Y. J., & Chang, M. C. (2021). Prevalence of Sarcopenia Among the Elderly in Korea: A Meta-Analysis. *Journal of preventive medicine and public health*, *54*(2), 96-102.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., & Bueno, R. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS genetics*, *5*(8), e1000602.

Chung, H., Jo, Y., Ryu, D., Jeong, C., Choe, S. K., & Lee, J. (2021). Artificial-intelligence-driven discover y of prognostic biomarker for sarcopenia. *Journal of Cachexia, Sarcopenia and Muscle*, *12*(6), 2220-2230.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Crews, D., & Gore, A. C. (2011). Life imprints: living in a contaminated world. *Environmental Health Perspectives*, *119*(9), 1208-1210.

Cruz-Jentoft, A. J., Baeyens, J. P., Bauer, J. M., Boirie, Y., Cederholm, T., Landi, F., Martin, F. C., Michel, J.-P., Rolland, Y., & Schneider, S. M. (2010). Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People. *Age and Ageing*, *39*(4), 412-423.

Cruz-Jentoft, A. J., Baeyens, J. P., Bauer, J. M., Boirie, Y., Cederholm, T., Landi, F., Martin, F. C., Michel, J.-P., Rolland, Y., Schneider, S. M., Topinková, E., Vandewoude, M., & Zamboni, M. (2010). Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People. *Age and Ageing*, *39*(4), 412-423.

Day, K., Waite, L. L., Thalacker-Mercer, A., West, A., Bamman, M. M., Brooks, J. D., Myers, R. M., & Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology*, *14*(9), R102.

Doshi-Velez, F., Ge, Y., & Kohane, I. (2014). Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, *133*(1), e54-e63.

Esteller, M. (2006). The necessity of a human epigenome project. *Carcinogenesis*, *27*(6), 1121-1125.

Evans, W. J. (1995). What is sarcopenia? *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 50(Special_Issue), 5-8.*

Feil, R., & Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, *13*(2), 97-109.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, *447*(7143), 433-440.

Fortin, J.-P., Timothy J. Triche, J., & Hansen, K. D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, 33(4), 558-560.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, *13*(2), 135-145.

Godfrey, K. M., & Barker, D. J. (2001). Fetal programming and adult health. *Public health nutrition*, *4*(2b), 611-624.

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83-90.

Gu, S., Wang, L., Han, R., Liu, X., Wang, Y., Chen, T., & Zheng, Z. (2023). Detection of sarcopenia using deep learning-based artificial intelligence body part measure system (AIBMS). *Frontiers in Physiology*, *14*, 46.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389-422.

Hales, C. N., & Barker, D. J. (2001). The thrifty phenotype hypothesis: Type 2 diabetes. *British medical bulletin*, *60*(1), 5-20.

Hardy, T. M., & Tollefsbol, T. O. (2011). Epigenetic diet: impact on the epigenome and cancer. *Epigenomics*, *3*(4), 503-518.

Hartman, M., Loy, E. Y., Ku, C. S., & Chia, K. S. (2010). Molecular epidemiology and its current clinical use in cancer management. *The lancet oncology*, *11*(4), 383-390.

He, L., Khanal, P., Morse, C. I., Williams, A., & Thomis, M. (2019). Differentially methylated gene patterns between age-matched sarcopenic and non-sarcopenic women. *Journal of Cachexia, Sarcopenia and Muscle*, *10*(6), 1295-1306.

Health Examinees Study, G. (2015). The Health Examinees (HEXA) study: rationale, study design and baseline characteristics. *Asian Pacific* *journal of cancer prevention: APJCP, 16*(4), 1591-1597.

Hendriksen, J. M., Geersing, G. J., Moons, K. G., & de Groot, J. A. (2013). Diagnostic and prognostic prediction models. Journal of thrombosis and haemostasis : JTH, 11 Suppl 1, 129–141.

Ho, S.-M. (2010). Environmental epigenetics of asthma: an update. *Journal of Allergy and Clinical Immunology*, *126*(3), 453-465.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, *13*, 1-16.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics*, *15*(1), 41-51.

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., & Rieck, B. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, *26*(3), 364-373.

Ibrahim, K., May, C., Patel, H. P., Baxter, M., Sayer, A. A., & Roberts, H. (2016). A feasibility study of implementing grip strength measurement into routine hospital practice (GRImP): study protocol. *Pilot Feasibility Studies*, *2*, 27.

Izzo, A., Massimino, E., Riccardi, G., & Della Pepa, G. (2021). A narrative review on sarcopenia in type 2 diabetes mellitus: prevalence and associated factors. *Nutrients*, *13*(1), 183.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature genetics*, *33(3)*, 245-254.

Jang, H. C. (2018). How to diagnose sarcopenia in Korean older adults? *Annals of geriatric medicine and research*, *22*(2), 73.

Jeziorska, D. M., Murray, R. J., De Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., Chen, T., Li, E., Telenius, J., & Lynch, M. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proceedings of the National Academy of Sciences*, *114*(36), E7526-E7535.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118-127.

Jones, M. J., Goodman, S. J., & Kobor, M. S. (2015). DNA methylation and healthy human aging. *Aging cell*, *14*(6), 924-932.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484-492.

Kang, Y. J., Yoo, J. I., & Ha, Y. C. (2019). Sarcopenia feature selection and risk prediction using machine learning: A cross-sectional study. *Medicine (Baltimore)*, *98*(43).

Kanherkar, R. R., Bhatia-Dey, N., & Csoka, A. B. (2014). Epigenetics across the human lifespan. *Frontiers in cell and developmental biology, 2, 49.*

Kim, K. M., Lim, S., Choi, K. M., Kim, J. H., Yu, S. H., Kim, T. N., Song, W., Lim, J.-Y., Won, C. W., & Yoo, H. J. (2015). Sarcopenia in Korea: prevalence and clinical aspects. *노인병*.

Kim, M., & Won, C. W. (2020). Sarcopenia in Korean community-dwelling adults aged 70 years and older: application of screening and diagnostic tools from the Asian Working Group for Sarcopenia 2019 update. *Journal of the American Medical Directors Association*, *21*(6), 752-758.

Kim, M., Won, C. W., & Kim, M. (2018). Muscular grip strength normative values for a Korean population from the Korea National Health and Nutrition Examination Survey, 2014-2015. *PLoS One*, *13*(8), e0201275.

Kim, Y., Han, B.-G., & Group, K. (2017). Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. *International journal of epidemiology*, *46*(2), e20-e20.

Kim, Y. J. (2021). Machine Learning Models for Sarcopenia Identification Based on Radiomic Features of Muscles in Computed Tomography. *Int J Environ Res Public Health*, *18*(16).

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack Jr, C. R., Ashburner, J., & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, *131*(3), 681-689.

Kwak, J. Y., Hwang, H., Kim, S.-K., Choi, J. Y., Lee, S.-M., Bang, H., Kwon, E.-S., Lee, K.-P., Chung, S. G., & Kwon, K.-S. (2018). Prediction of sarcopenia using a combination of multiple serum biomarkers. *Scientific Reports*, *8*(1), 8574.

Kwon, H.-J., Ha, Y.-C., & Park, H.-M. (2016). Prevalence of sarcopenia in the Korean woman based on the Korean national health and nutritional examination surveys. *Journal of bone metabolism*, *23*(1), 23-26.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882-883.

Leong, D. P., Teo, K. K., Rangarajan, S., Lopez-Jaramillo, P., Avezum, A., Jr., Orlandini, A., Seron, P., Ahmed, S. H., Rosengren, A., Kelishadi, R., Rahman, O., Swaminathan, S., Iqbal, R., Gupta, R., Lear, S. A., Oguz, A., Yusoff, K., Zatonska, K., Chifamba, J., . . . Yusuf, S. (2015). Prognostic value of grip strength: findings from the Prospective Urban Rural Epidemiology (PURE) study. *The Lancet*, *386*(9990), 266-273.

Lim, S., Kim, J. H., Yoon, J. W., Kang, S. M., Choi, S. H., Park, Y. J., Kim, K. W., Lim, J. Y., Park, K. S., & Jang, H. C. (2010). Sarcopenic obesity: prevalence and association with metabolic syndrome in the Korean Longitudinal Study on Health and Aging (KLoSHA). *Diabetes Care*, *33*(7), 1652-1654.

Lin, S., Chen, C., Cai, X., Yang, F., & Fan, Y. (2022). Development and Verification of a Combined Diagnostic Model for Sarcopenia with Random Forest and Artificial Neural Network. Computational and Mathematical Methods in Medicine, 2022.

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T. K., Vilo, J., Salumets, A., & Tõnisson, N. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology*, *15*(4), 3248.

Lorincz, M. C., Dickerson, D. R., Schmitt, M., & Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature structural & molecular biology*, *11*(11), 1068-1075.

Marks A. R. (2023). Targeting ryanodine receptors to treat human diseases. The Journal of clinical investigation, 133(2), e162891.

Maksimovic, J., Gordon, L., & Oshlack, A. (2012). SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, *13*(6), R44.

Martin, E. M., & Fry, R. C. (2018). Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annual review of public health*, *39*, 309-333.

Marx, S. O., & Marks, A. R. (2013). Dysfunctional ryanodine receptors in the heart: new insights into complex cardiovascular diseases. Journal of molecular and cellular cardiology, 58, 225–231.

Massimino, E., Izzo, A., Riccardi, G., & Della Pepa, G. (2021). The impact of glucose-lowering drugs on sarcopenia in type 2 diabetes: current evidence and underlying mechanisms. *Cells*, *10*(8), 1958.

McKee, A., Morley, J. E., Matsumoto, A. M., & Vinik, A. (2017). Sarcopenia: an endocrine disorder? *Endocrine Practice*, *23*(9), 1143-1152.

Mesinovic, J., Zengin, A., De Courten, B., Ebeling, P. R., & Scott, D. (2019). Sarcopenia and type 2 diabetes mellitus: a bidirectional relationship. *Diabetes, metabolic syndrome and obesity: targets and therapy*, 1057-1072.

Migliavacca, E., Tay, S. K. H., Patel, H. P., Sonntag, T., Civiletto, G., McFarlane, C., Forrester, T., Barton, S. J., Leow, M. K., Antoun, E., Charpagne, A., Seng Chong, Y., Descombes, P., Feng, L., Francis-Emmanuel, P., Garratt, E. S., Giner, M. P., Green, C. O., Karaz, S., . . . Feige, J. N. (2019). Mitochondrial oxidative capacity and NAD(+) biosynthesis are reduced in human sarcopenia across ethnicities. *Nature communications,10*(1), 5808.

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*, *98*(9), 683-690.

Moore, L. D., Le, T., & Fan, G. (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, *38*(1), 23-38.

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., & Beck, S. (2014). ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, *30*(3), 428-430.

Mossotto, E., Ashton, J., Coelho, T., Beattie, R., MacArthur, B., & Ennis, S. (2017). Classification of paediatric inflammatory bowel disease using machine learning. *Scientific Reports*, *7*(1), 1-10.

Nitert, M. D., Dayeh, T., Volkov, P., Elgzyri, T., Hall, E., Nilsson, E., Yang, B. T., Lang, S., Parikh, H., Wessman, Y., Weishaupt, H., Attema, J., Abels, M., Wierup, N., Almgren, P., Jansson, P. A., Rönn, T., Hansson, O., Eriksson, K. F., . . . Ling, C. (2012). Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. *Diabetes*, *61*(12), 3322-3332.

Patel, S. J., Chamberlain, D. B., & Chamberlain, J. M. (2018). A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. *Academic emergency medicine*, *25*(12), 1463-1470.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Priori, S. G., & Napolitano, C. (2005). Cardiac and skeletal muscle disorders caused by mutations in the intracellular Ca 2+ release channels. *The Journal of clinical investigation*, *115*(8), 2033-2038.

Rao, X., Evans, J., Chae, H., Pilrose, J., Kim, S., Yan, P., Huang, R., Lai, H., Lin, H., & Liu, Y. (2013). CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, *32*(38), 4519-4528.

Rossi, A. P., Fantin, F., Micciolo, R., Bertocchi, M., Bertassello, P., Zanandrea, V., Zivelonghi, A., Bissoli, L., & Zamboni, M. (2014). Identifying sarcopenia in acute care setting patients. *Journal of the American Medical Directors Association 15, no. 4 (2014): 303-e7.*

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

Santilli, V., Bernetti, A., Mangone, M., & Paoloni, M. (2014). Clinical definition of sarcopenia. *Clinical cases in mineral and bone metabolism*, *11*(3), 177.

Santos, A. G., da Rocha, G. O., & de Andrade, J. B. (2019). Occurrence of the potent mutagens 2- nitrobenzanthrone and 3-nitrobenzanthrone in fine airborne particles. *Scientific reports, 9(1), 1.*

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Frontiers in aging neuroscience*, *9*, 329.

Seaborne, R. A., Strauss, J., Cocks, M., Shepherd, S., O'Brien, T. D., van Someren, K. A., Bell, P. G., Murgatroyd, C., Morton, J. P., Stewart, C. E., & Sharples, A. P. (2018). Human Skeletal Muscle Possesses an Epigenetic Memory of Hypertrophy. *Scientific reports*, *8*(1), 1898.

Seok, M., & Kim, W. (2023). Sarcopenia Prediction for Elderly People Using Machine Learning: A Case Study on Physical Activity. *Healthcare*, *11*(9), 1334.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., & Campbell, H.

(2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, *89*(5), 607-618.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology, 3(1).*

Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, *16*(4), 253-259.

Tan, M. H. (2020). *Diabetes Mellitus: Impact on Bone, Dental and Musculoskeletal Health*. Academic Press.

Tang, W.-y., Morey, L. M., Cheung, Y. Y., Birch, L., Prins, G. S., & Ho, S.-m. (2012). Neonatal exposure to Estradiol/Bisphenol A alters promoter methylation and expression of Nsbp1 and Hpcal1 genes and transcriptional programs of Dnmt3a/b and Mbd2/4 in the RatProstate gland throughout life. *Endocrinology*, *153*(1), 42-55.

Tang, W.-Y., Newbold, R., Mardilovich, K., Jefferson, W., Cheng, R. Y., Medvedovic, M., & Ho, S.-M. (2008). Persistent hypomethylation in the promoter of nucleosomal binding protein 1 (Nsbp 1) correlates with overexpression of Nsbp 1 in mouse uteri neonatally exposed to diethylstilbestrol or genistein. *Endocrinology*, *149*(12), 5922-5931.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., & Jacobs, I. J. (2009). An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*, *4*(12), e8274.

Turner, D. C., Gorski, P. P., Maasar, M. F., Seaborne, R. A., Baumert, P., Brown, A. D., Kitchen, M. O., Erskine, R. M., Dos-Remedios, I., Voisin, S., Eynon, N., Sultanov, R. I., Borisov, O. V., Larin, A. K., Semenova, E. A., Popov, D. V., Generozov, E. V., Stewart, C. E., Drust, B., . . . Sharples, A. P. (2020). DNA methylation across the genome in aged human skeletal muscle tissue and muscle-derived cells: the role of HOX genes and physical activity. *Scientific reports, 10(1), 15360.*

von Haehling, S., Morley, J. E., & Anker, S. D. (2010). An overview of sarcopenia: facts and numbers on prevalence and clinical impact. *Journal of Cachexia, Sarcopenia and Muscle*, *1*, 129-133.

Waljee, A. K., Higgins, P. D., & Singal, A. G. (2014). A primer on predictive models. Clinical and translational gastroenterology, 5(1), e44.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. *Encyclopedia of machine learning*, *15*(1), 713-714.

Widschwendter, M., Apostolidou, S., Raum, E., Rothenbacher, D., Fiegl, H., Menon, U., Stegmaier, C., Jacobs, I. J., & Brenner, H. (2008). Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study. *PLoS One*, *3*(7), e2656.

Woodson, K., Mason, J., Choi, S.-W., Hartman, T., Tangrea, J., Virtamo, J., Taylor, P. R., & Albanes, D. (2001). Hypomethylation of p53 in peripheral blood DNA is associated with the development of lung cancer. *Cancer Epidemiology Biomarkers & Prevention*, *10*(1), 69-74.

Xu, W., Wang, M., Jiang, C.-M., & Zhang, Y.-M. (2011). Anthropometric equation for estimation of appendicular skeletal muscle mass in Chinese adults. *Asia Pacific journal of clinical nutrition*, *20*(4), 551-556.

Yu, J., Shin, M., & Kwon, T. (2017). Analysis of research trend on machine learning based malware mutant identification. *Review of KIISC*, *27*(3), 12-19.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). 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.

Zhang, X., & Ho, S.-M. (2011). Epigenetics meets endocrinology. *Journal of molecular endocrinology*, *46*(1), R11-R32.

Zhou, W., Laird, P. W., & Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic acids research, 45(4), e22-e22.*

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, *50*, 71-91.

Zupo, R., Moroni, A., Castellana, F., Gasparri, C., Catino, F., Lampignano, L., Perna, S., Clodoveo, M. L., Sardone, R., & Rondanelli, M. (2023). A Machine-Learning Approach to Target Clinical and Biological Features Associated with Sarcopenia: Findings from Northern and Southern Italian Aging Populations. *Metabolites*, *13*(4), 565.

Zykovich, A., Hubbard, A., Flynn, J. M., Tarnopolsky, M., Fraga, M. F., Kerksick, C., Ogborn, D., MacNeil, L., Mooney, S. D., & Melov, S. (2014). Genome-wide DNA methylation changes with age in disease-free human skeletal muscle. Aging cell, 13(2), 360–366. https://doi.org/10.1111/acel.12180

Chpater 7. 국문 초록

노인 인구 비율이 증가하면서 노인성 질환에 관심도가 높아지고 있다. 근감소증(sarcopenia)이 그중 하나이다. 근감소증(sarcopenia)은 신체적 노쇠와 비슷한 개념으로 노화에 따른 근육량, 근력의 감소와 신체기능의 감소가 이루어지는 상태를 의미한다. 이는 사망률 증가, 골다공증, 골절, 각종 질환이 동반된다. 과거에는 자연스러운 노화의 한 과정으로 여겼지만, 최근 세계보건기구(WHO)에서는 근감소증을 정식 질병 코드로 등재했으며, 한국 또한 근감소증 질병코드(M62.5)를 부여했다.

노화로 인해 근육량, 근력 및 신체 기능이 감소하면서 이차적인 질병 유발 가능성이 커진다. 이러한 근육량 및 기능의 감소 속도는 개인마다 차이가 있다. 노화에 따른 근육의 변화는 고정된 유전적 요인에 기인할 수 있지만, 환경적인 요인과 유전적 요인과의 상호작용으로도 발생한다. 따라서 근감소증 후성유전학적 연구가 필요하지만 이에 대한 연구는 아직 부족한 실정이다.

후성유전학은 DNA 염기서열의 변화 없이 유전자 발현에 영향을 미치는 현상이며, 이는 생애 동안 접하는 환경적 요인에 의해 영향을 받는 것으로 알려져 있다. DNA 메틸화와 히스톤 변형은 후성유전학의 주요 메커니즘으로, DNA 메틸화는 노화와 만성 질환의 발병에 중요한 역할을 한다. 노화 및 다양한 질병으로 인해 근육의 양과 기능의 변화가 DNA 메틸화와 관련이 있다는 점을 감안할 때 근감소증과 DNA 메틸화 사이의 잠재적인 관계가 제안되었다. 따라서 본 연구의 목적은 한국인유전체역학조사(KoGES) 데이터를 활용해 한국인의 근감소증 잠재적 바이오마커를 발굴하고 해당 유전체를 이용하여 근감소증 진단 및 예측 모델을 개발하고자 한다.

2004년부터 2013년까지 Korea Genome and Epidemiology Study (KoGES)의 데이터를 활용하였다. 총 110명 (남성: 82명, 여성: 28명)의 차등 메틸화 DNA probe 를 분석하였다. 피험자는 근육량 (사지골격근량, appendicular skeletal muscle index; ASMI)과 근력 (악력, handgrip)의 두 변수를 기준을 이용하였다. 본 연구에서는 두 변수 데이터의 상위, 하위 25%로 나누어 근감소증 그룹을 결정하였다. 메틸화 데이터는 Infinium 사의 Infinium Methylation Epic Beadchip 로 어세이 된 자료들은 DNA 메틸화 배치 효과에 대한 정규화 및 보정 등 적절한 데이터 처리 단계를 이용한 후 유전자 내 총 740,000개 이상의 마커를 얻을 수 있었다. 이후 $|\mathbf{logFC}| > \mathbf{0.15}$ 그리고 $\mathbf{FDR\ adjusted}\ p - \mathbf{value} < \mathbf{0.05}$를 기준으로 두어 차등 메틸화 DNA probe 를 분석하였다. 남성은 과메틸화 99 probe, 저메틸화 67 probe 가 발견되었으나 여성의 경우에는 임곗값을 충족하지 못하여 차등 메틸화 분석을 수행할 수 없었다. 따라서 여성 그룹에 대한 데이터는 차등 메틸화와 관련하여 유의미한 결과가 부족하여 제외되었다.

주요한 바이오마커를 식별하기 위해 남성의 과메틸화 99 probe 와 저메틸화 67를 합쳐 총 166개 probe 를 분석하였다. 변수 데이터가 정규 분포를 이룰 수 있도록 피어슨 상관계수 (Pearson correlation)를 사용하여 134개 probe 이 제거되었다. 134개 probe 중 유의미한 변수를 선택하기 위해 재귀적 특성 제거 교차 검증 (recursive feature elimination cross-validation; RFECV)을 사용하였다. 최종적으로 유의미한 연관성을 가진 총 10개가 확인되었다. 확인된 probe 은 majority voting 앙상블을 이용하여 근감소증 진단 모델을 구축하였다. 사용된 앙상블 기법은 모델 성능을 개선할 수 있는 기법으로, 단일 모델보다 더 나은 성능을 달성할 수 있기 때문에 사용되었다. Train 과

test 데이터 세트는 7:3으로 나누어 분석하였다. Train 데이터 세트는 Decision tree, random forest, logistic regression, K-Nearest Neighbors, Naïve Bayes 4가지 알고리즘을 이용하여 학습되었으며 사용된 개별 모델의 예측을 결합하여 majority voting 값을 도출하였다. 마지막으로 test 데이터 세트를 이용하여 진단 모델을 평가하고, area under the curve (AUC) 값으로 진단 성능을 평가하였다. 구축된 모델은 한국인 근감소증 유전자 표현형을 식별하는 데 있어 높은 정확도(96%)를 보였다. 또한 10개의 probe 중 TCF12, RYR2, 그리고 ZNF415는 근감소증과 관련된 유전체로 확인되었다. TCF12 유전은 근육 발달 및 재생에 영향을 주며, RYR2는 심장 근육과 관련된 유전체이지만 근육과 관련된 RYR1 유전체와 같은 receptor 에서 방출되기 때문에 RYR1과 함께 근육에 영향을 줄 수 있다. 마지막으로 ZNF415는 다른 유전자에 비해 광범위하게 연구되고 특성화되지는 않았지만, 유전자 조절 및 전사 조절을 포함한 다양한 세포 과정에 관여한다. 선행 연구에서 ZNF415 유전체가 근감소증과 연관이 있는 biomarker 로 확인되어 근육에 영향을 주는 유전체라고 시사할 수 있다.

노화에 따라 근육량, 근력의 감소 그리고 신체의 기능의 감소가 야기되기 때문에 확인된 10개의 probe 을 이용하여 연령에 따른 발현율의 차이를 확인하였다. 그 중 RYR2는 나이와 가장 큰 음의 상관관계 (r = -0.64)를 보여 이를 통해서 근감소증은 노인에게서만 나타나는 질환이 아니라 중장년층에서도 발현될 수 있기 때문에 근감소증은 초기에 예방을 해야 한다고 시사할 수 있다.

본 연구는 머신러닝 기법을 통해 근감소증의 잠재적 유전자를 확인하였으며 근감소증 진단 성능이 높은 초기 예측 및 진단 모델을 개발하였다. 한국인의 근감소증 예측력이 향상되었지만, 확인된 메틸화

probe 와 근감소증과의 관계를 규명하기 위해서는 세포 및 분자생물학적 검증을 종합적으로 수행하는 것이 중요하다. 본 연구를 통해 도출된 근감소증의 잠재적 유전자는 한국인의 근감소증 위험의 근본적인 기전에 대한 귀중한 통찰력을 제공하고, 표적 중재를 위한 유용한 참고 자료가 될 것이다.

Keywords: 근감소증, 후성유전, DNA 차등 메틸화, 머신러닝, 한국인유전체역학조사사업

Student Number: 2021-28886